BREAST CANCER IMAGE CLASSIFICATION USING EXTERNAL ATTENTION MULTILAYER PERCEPTRON BASED TRANSFORMER

By

Shemonti Barua

1708008

A thesis submitted in partial fulfilment of the requirements for the degree of BACHELOR of SCIENCE in ELECTRONICS AND TELECOMMUNICATION ENGINEERING

Department of Electronics and Telecommunication Engineering

CHITTAGONG UNIVERSITY OF ENGINEERING AND TECHNOLOGY

APRIL 2023

Declaration

I hereby declare that the work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the Thesis contains no material previously published or written by another person except where due reference is cited. Furthermore, the Thesis complies with PLAGIARISM and ACADEMIC INTEGRITY regulation of CUET.

Shemonti Barua
1708008
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology (CUET)

Approval by the Supervisor(s)

This is to certify that Shemonti Barua has carried out this research work under my supervision, and that she has fulfilled the relevant Academic Ordinance of the Chittagong University of Engineering and Technology, so that she is qualified to submit the following Thesis in the application for the degree of BACHELOR of SCIENCE in Electronics and Telecommunication Engineering. Furthermore, the Thesis complies with the PLAGIARISM and ACADEMIC INTEGRITY regulation of CUET.

Dr Md Saiful Islam
Associate Professor
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology

Acknowledgement

First and foremost, I want to thank God for helping me with my studies and for protecting, blessing, and guiding me. Without my faith, I would never have been able to accomplish this.

Second, I want to sincerely thank my supervisor, Associate Professor Dr MD Saiful Islam sir for all of his help, patience, knowledge, and motivation throughout the years. The success of this research is primarily attributable to his distinctive personality as a friend and supervisor. Without my supervisor's expert direction and support, the research objectives would not have been met.

Last but not least, I would want to thank my wonderful family for their steadfast support and love as I pursued my studies. I appreciate your unwavering support and your encouragement to persevere when things get difficult. I count it a blessing that I have come across family.

.

Abstract

Breast cancer is now the most prevalent illness diagnosed today and main factor in all cancer-related deaths among women. Due to this, many scholars in the field of healthcare are interested in the classification of breast cancer. The use of various deep learning techniques and self-attention-based transformers for breast image analysis has increased during the last several years. But self-attention is quadratic in complexity and disregards the possibility of sample to sample association. The self-attentional capacity and interpretability are constrained by this quadratic complexity. In order to address the issue, this study presents a novel attention mechanism based on two external, tiny, learnable, shared retention units that we refer to as the external attention multi layer perceptron (EAMLP). Because external attention has a linear complexity, this suggested approach uses it to reduce operational complexity. External attention also effectively takes into consideration the links between all of the datapoints. The effectiveness of the study is assessed based on its accuracy, sensitivity, and specificity. In this work two publicly available datasets has been used among them one dataset is used for both training and testing and another is used for testing purpose. The proposed model is evaluated by using two different image size for evaluating the model performance. A self-attention based transformer has also experimentally tested to compare with the proposed model. The results reveal that the proposed model obtained the highest accuracy, sensitivity, specificity of 95.73%, 96.11%, and 95.15% respectively.

বিমূর্ত

স্তন ক্যান্সার এখন সবচেয়ে ঘন ঘন নির্ণয় করা ক্যান্সার এবং মহিলাদের মধ্যে ক্যান্সার মৃত্যুর বিশ্বব্যাপী কারণ এই কারণে, স্বাস্থ্যসেবা ক্ষেত্রের অনেক পণ্ডিত স্তন ক্যান্সারের শ্রেণীবিভাগে আগ্রহী। বিগত কয়েক বছর ধরে, স্তনের চিত্র বিশ্লেষণের জন্য বিভিন্ন গভীর শিক্ষা পদ্ধতি এবং স্ব-অনটেনশন ভিত্তিক ট্রান্সফরমার ব্যবহারে বৃদ্ধি পেয়েছে। কিন্তু স্ব-মনোযোগ জটিলতায় চতুর্মুখী এবং নমুনা থেকে নমুনা সংযোগের সম্ভাবনাকে উপেক্ষা করে। এই দ্বিঘাত জটিলতা স্ব-মনোযোগের ব্যাখ্যা এবং ক্ষমতাকে সীমিত করে।সমস্যা সমাধানের জন্য, এই কাজটি দুটি বাহ্যিক, ছোট, শেখার যোগ্য, শেয়ার্ড রিটেনশন ইউনিটের উপর ভিত্তি করে একটি অভিনব মনোযোগ পদ্ধতির প্রস্তাব করে যাকে আমরা বাহ্যিক মনোযোগ মাল্টি লেয়ার পারসেপ্টরন (EAMLP) বলি। বাহ্যিক মনোযোগের রৈখিক জটিলতা থাকায় এই প্রস্তাবিত মডেলটি অপারেশনের জটিলতা কমানোর জন্য বাহ্যিক মনোযোগ নিযুক্ত করে। সমস্ত ডেটাপয়েন্টের মধ্যে আন্তঃসংযোগগুলিও বাহ্যিক মনোযোগের দ্বারা কার্যকরভাবে বিবেচনা করা হয়।অধ্যয়নের কর্মক্ষমতা নির্ভুলতা, সংবেদনশীলতা, নির্দিষ্টতার সাথে পরিমাপ করা হয়। এই কাজে দুটি সর্বজনীনভাবে উপলব্ধ ডেটাসেট ব্যবহার করা হয়েছে তাদের মধ্যে একটি ডেটাসেট প্রশিক্ষণ এবং পরীক্ষা উভয়ের জন্য ব্যবহৃত হয় এবং অন্যটি পরীক্ষার উদ্দেশ্যে ব্যবহৃত হয়। প্রস্তাবিত মডেলটি মডেলের কার্যকারিতা মূল্যায়নের জন্য দুটি ভিন্ন চিত্রের আকার ব্যবহার করে মূল্যায়ন করা হয়। একটি স্ব-মনোযোগ ভিত্তিক ট্রান্সফরমারও প্রস্তাবিত মডেলের সাথে তুলনা করার জন্য পরীক্ষামূলকভাবে পরীক্ষা করেছে। ফলাফলগুলি প্রকাশ করে যে প্রস্তাবিত মডেলটি যথাক্রমে 95.73%, 96.11% এবং 95.15% এর সর্বোচ্চ নির্ভুলতা, সংবেদনশীলতা, নির্দিষ্টতা পেয়েছে।

Table of Contents

List of Figures

Fig. No. **Figure Caption** **Page No.**

List of Tables

Table No. **Table Caption** **Page No.**

Chapter 1: INTRODUCTION

This chapter includes the background study behind the cancer, breast cancer respectively. Moreover this chapter describes the problem statement of breast cancer classification techniques. It illustrates the need of diagnosis of breast cancer by using different techniques. It also depicts the scope of study of breast cancer which includes different machine learning approaches. Also this section presents some recent works relates to breast cancer image classification. Furthermore, it narrates the objective and contribution of this research work. In addition it also delineates the study of significance on this research subject. In last section it illustrates the thesis layout which describes the whole summary point of this research work.

1.1 RESEARCH BACKGROUND

Cancer is a set of disorders that originate when aberrant cells grow out of control or invade other bodily regions. Cancer may originate basically anywhere in the millions of cells that make up the human body. As the body requires new cells, human cells regularly divide to create them. Tissue replace old counterparts when they die

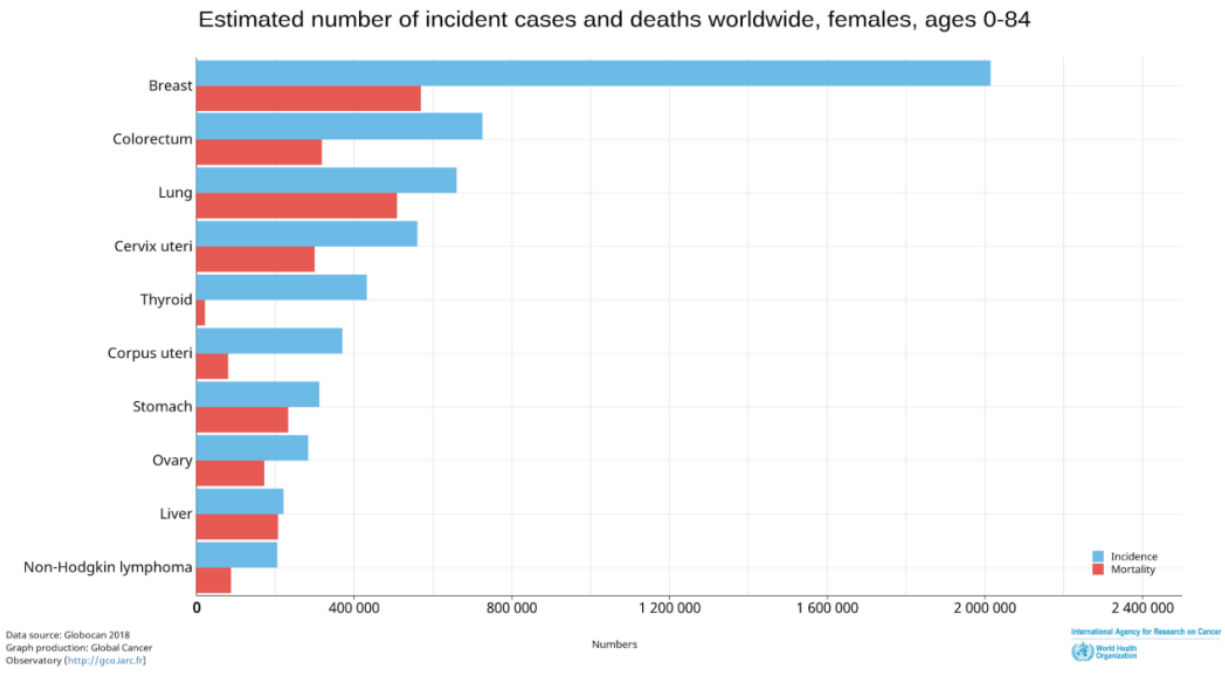


Figure.1.1 Common women cancers[1].

prematurely of age or injury. Cancerous tumors may invade neighboring tissues, spread to far-off tissues throughout the body, or do both. Cancerous tumors are also characterized as malignant tumors. Leukemias and other blood cancers seldom evolve to solid tumors, but many other malignancies do. Noncancerous tumors do not invade or expand to surrounding tissues. While benign tumors ordinarily don't, malignant ones occasionally do after excision. Yet, benign tumors may develop into exceedingly large masses. Some of these, like benign brain tumors, can have deadly adverse effects[1]. In 2020, there will likely be 19.3 million more deaths from cancer and over 10 million cancer-related deaths worldwide[2].

Breast cancer and other cancers are the worst of them all and the one that strikes women the most frequently. With an anticipated 2.3 million new cases, female breast cancer has overtaken lung cancer as the most prevalent malignancy diagnosed. Early diagnosis and rapid chemotherapy for breast cancer can help lower the death rate by allowing early detection and accurate cancer type identification.

Breast cancer has already taken 500,000 lives and there are already more than 1.7 million new cases reported each year. These numbers are expected to significantly increase in the next years. Hence, early identification of invasive ductal carcinoma is essential in order to offer breast cancer patients with proper therapy, lower the devastation and mortality rate, and increase survival rates.

One of the most frequent causes of mortality globally is cancer. Unfortunately, cancer is an illness that spreads across cells and worsens daily. According to the World Health Organization (WHO), cancer was the second-leading reason for death globally in 2018, accounting for 9.6 million deaths. Lung cancer, prostate cancer, skin cancer, and breast cancer are only a few of the numerous types of cancer. According to the International Agency for Research on Cancer, breast cancer is one of the diseases that affect women the most frequently. Fig 1 depicts the common women cancer also presents the increasing rate of breast cancer compared with other cancer. Moreover it depicts the death rate of different type of women cancers.

1.2 PROBLEM STATEMENT

Breast carcinoma is a prevalent disease that infects many women throughout their lifetimes, but it may also affect men. According to the Breast Cancer Institute [3], breast cancer is one of the diseases that kill the most women in the globe. Early detection is the finest and most efficient method of effectively managing cancer. On the other hand, delaying a diagnosis might lead to disease spreading throughout the body, making it harder to treat and manage. In addition, a late diagnosis reduces the possibility of effective treatment.

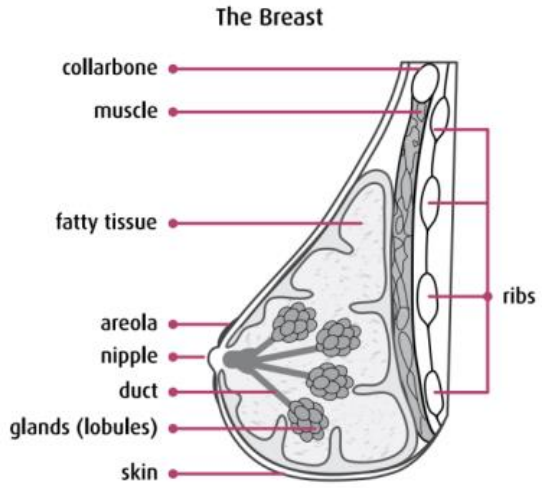


Figure.1.2 Common Structure of Breast[2].

See Figure. 1.2 for a breakdown of the many parts of the breast, including the connective tissue, fat, glands, and ducts, all of which are susceptible to cancer. Self-examination at home, breast screening, and doctor visits are a few of the many ways to find breast cancer early. These methods will reduce the chance of mortality and raise the possibility of a successful outcome. The most popular breast imaging methods are breast ultrasound, MRI, imaging techniques, thermography, mammography, cytopathological imaging, and histopathological imaging. Ultrasound and mammography are the two most often utilized early breast cancer diagnostic techniques. The most prevalent subtype of breast cancer (IDC) is invasive ductal carcinoma . Invasive ductal carcinomas account for around 80% of all cases of breast cancer [4]. Often, tissue examination slides are used by pathologists to identify this. On tissue

slides, pathologists typically concentrate on the area where invasive ductal carcinomas are present to assess the severity of this kind of cancer [5].

In order to differentiate between tissue areas associated with invasive ductal carcinomas and non-invasive ductal carcinomas, histological study must first make this distinction. The popular Bloom-Richardson and Nottingham grading procedures can be used to analyze tumor aggressiveness in further detail when IDC is found in the tissue slides. Invasive ductal carcinoma breast cancer identification is typically difficult since the pathologist must examine a significant area of healthy cells before describing the IDC. Yet, due to a variety of human factors, like interpretation mistakes brought on by fatigue or complex patterns that are tricky for the human eye to perceive, the pathologist's decision-making task may be difficult. In this way, early detection is necessary to provide breast cancer patients with appropriate care and to lower the morbidity and mortality rate.

1.3 SCOPE OF STUDY

Several experts and researchers in the field of health have used various techniques and conducted research to help diagnose breast cancer A Computer Aided Diagnosis (CAD) system that uses convolutional neural networks was used in a study to apply deep learning techniques to distinct breast cancer datasets. Particularly effective have been Deep Learning models and Convolutional neural networks (CNNs). In learning tasks, deep learning algorithms are so effective that they surpass the conventional approaches, such as feature extraction and machine learning [6–10].

CNNs and Deep Learning models have demonstrated potential in a variety of medical imaging applications for the detection of breast cancer during the past several years [11, 12]. In an effort to apply deep learning to image data sets, the majority of the images use complex model architecture made up of Alex Net [13] to VGG-16 [14], ResNet [15], Inception-V3 [16], and DenseNet [17] with functional deceptions like dropout regularization, batch normalization, transfer learning, and zero-shot training . Despite the fact that this enhances performance, deep networks and massive inputs come with significant computational costs and parameters,

which lengthen the time required for network training and tuning. So for this problem new mechanism has been introduced in recent times called attention mechanism. Neural networks' attention mechanisms have a propensity to resemble humans' cognitive attention. The prime purpose of this function is to attempt to highlight important information while underscoring unnecessary material. Researchers have looked at a variety of techniques leveraging this learning algorithm on breast histology pictures in recent years. To extract the input picture characteristics in Latent space for image classification, Ref. [18] applied an attention technique. Ref. [19] built a soft attention system into its structure to focus just on the area of interest. They used an attention method [20] and an adaptive spectral composition for classification. Using high-resolution image data from the BreakHis and Bach datasets, Li et al. [21] employed a multiview attention-guided ensemble learning detection network.

Many hybrid attention processes have also been employed to categorize breast image[22][23] created a decision model (DeNet) with a soft-attention classification network (SaNet) for the classification of breast photos. In order to automatically and reliably separate breast tumors from ultrasound images, [24] developed an adaptive attention U-net (AAU-net). [25] released a self-attention GRU model to detect invasive ductal carcinoma. In the suggested et al.[26] model, a number of self-attention heads extract spatial information, while an ensemble model (Densenet201 and VGG16) serves as the network's backbone for a more general feature extraction of the input images (regions of interest). According to et al. [37], a Multi-Task Learning Network with Context-Oriented Self-Attention (MTL-COSA) module could autonomously divide tumors into benign and malignant types A novel AI-based computer-aided diagnostic (CAD) framework called ETECADx is proposed[28]. It combines the advantages of ensemble transfer learning of convolutional neural networks with the self-attention mechanism of the vision transformer encoder (ViT).

In order to improve prediction accuracy, et al.[29] introduced a breast cancer diagnosis model based on Phrase level self-attention mechanisms that employs Phrase level context technology to train the model for reading medical reports in the same way as a healthcare professional with original, local, and global method context. Pictures of breast cancer were categorized using a channel attention module

with non-dimensionality reduction by Zou et al. [30]. [31] A thick remnant dual-shuffle attention network was employed by Chattopadhyay et al. to classify images of breast malignancy. A new breast cancer detection model called SAFNet is proposed in et al[32] based on ultrasound images and deep learning, with the backbone model being a pre-trained ResNet-18 combined with the spatial attention mechanism. In the supervised attention technique proposed by et al.[33] for the segmentation of breast cancer histopathology images using CNN, the regions of interest (RoI) are localized and used to guide the attention of the classification network simultaneously. The supervised attention mechanism, which has been postulated, suppresses activations in irrelevant and noisy regions while selectively activating neurons in diagnostically significant regions.

A multi-task learning (SHA-MTL) model is provided in [34] for the simultaneous segmentation and binary classification of breast ultrasound (BUS) images. This model uses soft and hard attention methods. et al. [35] created a spatial attention method based on human knowledge to train deep learning models informative tissue areas of interest. Our prediction algorithm for prognostic tissue areas is guided by the ensuing comprehensive attention information from the picture triplets. But Self-attention requires a lot of computing resources, particularly for longer sequences. It is difficult to manage extremely large sequences because of its quadratic complexity with respect to sequence length. Understanding the model's decision-making process might be tricky since self-attention techniques can be tough to comprehend.

1.4 OBJECTIVE

In order to resolve the aforementioned problems, this research suggests a lightweight model named external attention multi layer perceptron(EAMLP) based transformer. This research employs an attention model that consist of two retention unit. As the two retention unit are built using linear layers, they may be optimized end-to-end via back-propagation. They are constant across the whole dataset and unaffected by particular samples, which improves the regularization function and broadens the scope for generalization of the attention mechanism. We expand external attention's ability to learn various attentional properties for the same input by introducing the multi-head mechanism.

The key to external attention's small weight is that there are many less components in the retention unit than there are in the input feature, which results in a computational cost that is proportional to the number of input elements. The external retention units are designed to gather the most distinct traits throughout the whole dataset, gathering the most useful bits, and excluding competing information from other samples. Better performance on some tasks may result from the model learning more intricate and non-linear correlations between the input characteristics thanks to the MLP. As it can identify more abstract characteristics and patterns, the MLP can aid in the model's ability to generalize to fresh and unexplored data. The MLP can add stability to the model by preventing the attention mechanism from becoming too sensitive to noisy or irrelevant input features.

The MLP can be used to customize the attention mechanism to suit the specific needs of a given task, such as by introducing additional constraints or regularization. For performance monitoring, this suggested model utilised two different pixel sizes, one being 64x64 and the other being 112x112. For the breast image classification job, the suggested model, dubbed EAMLP, is contrasted with CNNs and the original transformer This is a summary of this work's contribution:

- To implement External attention, a cutting-edge attentional mechanism multi-layer perceptron(**EAMLP**) based transformer to take the role of self-attention in existing designs by reducing the complexity to O(n) for the breast images classification.
- To introduce two linear, small retention units for the proposed model through two linear layers and two normalization layers.
- To compare the model performance, specificity, sensitivity between state-of-the-art architectures like original transformers.
- To observe the best performance of the proposed model for different number of multi layer perceptron has been employed on the proposed model for two different pixel size.

1.5 STUDY SIGNIFICANCE

Breast cancer is the second leading cause of death for women. Early detection of breast cancer is the most accurate and practical way to treat the disease. As contrasted to other cancer forms, BC has a very high fatality rate. Imaging techniques include diagnostic mammography (x-rays), magnetic resonance imaging, ultrasound (sonography), and

thermography can be used to detect and diagnose BC [36]. There has been much study on imaging for screening mammography for more than 40 years [37]. Regrettably, a biopsy is the only accurate way to determine whether cancer is truly present. The surgical (open) biopsy (SOB), vacuum-assisted, core needle biopsy, and fine needle aspiration techniques are the most frequently used biopsy techniques [38]. Tissue or cell samples are gathered during the process and preserved over a glass microscope slide for later staining and microscopic inspection. The criterion for identifying almost all cancer forms, including BC, is a diagnosis made from a histological picture [39], [40]. Pathologists use a visual inspection of histological specimens under a microscope for the purposes of grading and staging to arrive at the final BC diagnosis. Expert pathologists are needed for the intricate, drawn-out process of histopathological analysis, which can be affected by factors like exhaustion and a lack of focus. Computer-assisted diagnosis (CAD) is badly needed, as stated by Gurcan et al. [41], to relieve the burden on pathologists by filtering out plainly benign regions so that the specialists can focus on the more challenging-to-diagnose cases [42]. One of the software technologies called computer-aided detection or computer-aided diagnosis helps clinicians identify or diagnose cancer more promptly and minimizes mortality by employing medical picture analysis.

In order to find cancer cells or categorize photos, convolution neural networks have lately been used in medical image analysis to examine a lot of data. Despite Deep Convolutional Neural Network's effectiveness for a variety of image classification tasks, extracting specific information from biological images is difficult and time-consuming. The deep networks' high computational requirements, parameters, and massive inputs lengthen the time required for network training and tuning. Learning Deep-Networks with extremely high input dimensions requires a substantially bigger network structure, more hidden layers, and hardware memory.

Although Deep Convolutional Neural Network has demonstrated its efficacy for a variety of picture classification tasks, it is challenging and time-consuming to extract all of the information available in biomedical images. Deep networks have significant computing requirements, which add to the time needed for network tuning and training. The following challenges are involved in applying deep learning to categorize medical images. A deep learning algorithm with the highest performance cannot initially be created using the properties of medical photos. Second, the organizational framework and training methods of the present deep learning networks are less suitable for use with

medical imaging. This could facilitate understanding of the model. Yet paying attention is a useful tactic that helps explain our model's behaviour and sheds light on why it behaves the way it does. The long execution time and difficulty in parallelizing the self-attention approach employing transformer are its only drawbacks.

However, even after it has been simplified, employing self-attention has a significant disadvantage due to the tremendous computational complexity of O(dN2). The quadratic complexity of the input pixel count prevents the direct application of self-attention to visuals. To solve this issue, a lightweight focus on simplifying the model and enhancing performance should be implemented.

1.6 THESIS LAYOUT

In this work, chapter 2 is devoted to a review of the research on the identification and categorization of breast cancer utilizing various deep learning and attention-based transformer models. The primary focus of this dissertation is categorizing medical images, and the main goal of this chapter is to examine the relevant literature in that area.

In Chapter 3, the strategy for classifying breast cancer from histopathology pictures using an attention-based transformer model is discussed. Moreover, the structure and operation of proposed framework has also depicted in this chapter. Furthermore, this chapter provides the information of datasets and other traditional approaches which is important for this methodology.

Chapter 4 displays the experiment model outcomes and compares them to current best practices. Also, elaborate on the attention process and the remarkable accuracy attained. The comparison has been showed by using two section one is performance of the proposed model and another is comparison with other state of work.

The thesis conclusions are presented in chapter 5 along with discussion of additional research. In this chapter this conclusion has been divided into three portion. The limitation of this research work and future work of this research has been described in this chapter.

Chapter 2: LITERATURE REVIEW

The categorization of medical photographs is the primary concern of this dissertation, and the review of the pertinent literature in that field is the major objective of this chapter. We quickly go through how various picture and dataset types may be used to classify breast cancer. Many techniques for recognizing and classifying breast cancer have been developed and improved using deep neural networks with varied topologies. Classifying histopathology pictures using the cellular structure, intricate morphology, and texture is a challenging problem in medical imaging interpretation and classification. Deep learning models and pre-trained deep neural networks are the only two contemporary approaches that have been suggested to address the difficult problem of photo categorization.

2.1 TRADITIONAL APPROACHES TO MEDICAL IMAGE CLASSIFICATION

Deep learning is now generating a lot of attention in the healthcare industry. This is because, in addition to picture classification, deep learning methods, such convolutional neural networks (CNN), are now exceptionally good at carrying out all other kinds of visual identification tasks, such as object detection, semantic segmentation, and image classification. The development of Convolutional Neural Network Improvement for Breast Cancer Classification (CNNI-BCC) by [43] has the potential to tip the scales in favor of a more favorable and least invasive method of breast cancer screening. [44] We offer a method for dividing breast cancer into several subtypes using a recently created class structure-based deep convolutional neural network (CSDCNN).

An implementation of a convolutional Deep-Net Model based on the extraction of random patches and the enforcement of depthwise convolutions is provided in [45] for the training and classification of well-known benchmark Breast Cancer histopathology imagesPratiher et al[46] .'s focus The suggested method uses class-specific manifold learning (CSML), which is enabled by deep neural networks, together with a discriminative ensemble of local shallow signatures based on hashing of Histology pictures (DNNs). A DCNN descriptor and pooling operation are described as an effective deep learning-based solution in [47] for the categorization of breast cancer. Malon and Christopher[48] have demonstrated that convolutional neural networks

provide a versatile way for identifying regions of pathological importance in biopsy images. Cireşan, Dan C., et al. [49] employ a supervised Deep Neural Network (DNN) as a powerful pixel classifier. The DNN is a convolutional neural network with max-pooling (MP) (CNN). Le et altwo-layer .'s neural network was employed to quantitatively analyze tissue slices and pinpoint the glioblastoma multiforme cells that were necrotic, apoptotic, and alive (GBM). The tough task of recognizing mitosis in BCa images from histology has recently been been taken up by CNN models[51-53]. [54] presents an original CNN application for visual image processing in digital pathology and compares it to well-known handcrafted features. More than 1200 DP pictures were employed for assessment in [55], the biggest thorough research of DL techniques in DP to date.

2.2 DEEP LEARNING APPROACHES TO MEDICAL IMAGE CLASSIFICATION

Rakhlin et al. developed a sophisticated learning-based technique in [56] for classifying pictures of breast tissue stained with H&E. 20 crops of 400400 and 650650 pixels each were taken from each image. Lastly, feature extractors were created using the pre-trained ResNet-50, InceptionV3, and VGG-16 networks. Three-norm pooling was used to integrate the acquired features into a single feature vector. The gathered deep features were categorized using a 10-fold crossvalidation LightGBM classifier. The total average accuracy for classifying breast cancer histology images using this approach was 87.2 2.6%. Kwok [57] used four DCNN architectures—VGG19, InceptionV3, InceptionV4, and InceptionResnetV2—for both multi-class and binary classification while classifying H&E stained histological breast cancer pictures. The investigation uses 5600 patches from the images, each with a size of 14951495 pixels and a stride of 99 pixels.

To increase the method's accuracy, many data augmentation techniques were applied. InceptionResnetV2 has the highest level of accuracy in Kwok's study, with multi-class classification accuracy of 79% and binary classification accuracy of 91%. Vang et al. [58] published an ensemble-based InceptionV3 architecture for multi-class breast cancer picture categorization. Their proposed ensemble classifier utilized majority voting, gradient boosting machine (GBM), and logistic regression to obtain the final image-wise prediction. The Vahadane [59] stain-normalization technique, which had an

accuracy of 87.50% with additional improvement, was used to normalize the stain photographs. Nawaz et al. [60] used a fine-tuned AlexNet architecture to automatically classify breast cancer. It took an accuracy of 75.73% for the patch-wise dataset and 81.25% for the image-wise dataset to successfully extract the 512x512-pixel-sized patches from training images.

Although Deep Convolutional Neural Network has demonstrated its efficiency for a variety of picture classification tasks, it is challenging and time-consuming to extract all of the information contained in biomedical images. The extensive computer requirements, deep network-specific characteristics, and massive inputs increase the time required for network training and tuning. For deep learning with very wide input dimensions, a much bigger network structure with more hidden layers and hardware memory is needed. This problem is addressed with a deep-NET structure that leverages SeparableConv2D to guarantee depth-wise convolutions. In this construction, the dilated convolution is combined with the depth-wise separable convolution architecture.

2.3 DIFFERENT ATTENTION MECHANISM APPROACHES TO MEDICAL IMAGE CLASSIFICATION

Neural networks' attention mechanisms have a propensity to resemble humans' cognitive attention. There is a tendency for attentional systems to mirror cognitive attention in humans. The main objective of this function is to attempt to highlight important information while underscoring unnecessary material. In recent years, researchers have examined a number of methods that make use of this attention mechanism on images of breast histology. To extract the input picture characteristics in Latent space for image classification, Ref. [18] applied an attention technique. In order to concentrate just on the region of interest, Ref. [19] implemented a soft attention network into its architecture. For classification, they employed an attention strategy [20] and adaptive spectral composition.

High-resolution picture data was applied by Li et al. [21] to the BreakHis and Bach datasets using a multiple feature attention-guided instance based detection network. Many hybrid attention processes have also been employed to categorize breast image[22]. A decision network (DeNet) with a soft-attention classification model was developed by [23] for the categorization of breast pictures (SaNet). [24] created an

adaptive attention U-net to consistently and automatically distinguish breast cancers from ultrasound pictures (AAU-net). A self-attention GRU model to identify invasive ductal carcinoma was published by [25]. A number of self-attention heads recover spatial information in the suggested et al.[26] model, while a hybrid learning (Densenet201 and VGG16) serves as the network's backbone for a more thorough feature extraction of the input pictures . According to et al. [27], an MTL-COSA module (Multi-Task Learning Network utilizing Context-Oriented Self-Attention) can automatically distinguish between benign and malignant tumors. It is suggested to use ETECADx, an unique AI-based computer-aided diagnostic (CAD) framework[28]. It combines the advantages of convolutional neural network ensemble transfer learning with the self-attention technique of the vision transformer encoder (ViT). et al. [39] developed a phrase level self-attention mechanism-based breast cancer diagnosis model to increase the precision of their predictions.

By utilizing a channel attention unit with non-dimensionality removal, Zou et al. [30] classified images of breast cancer. [31] A thick residual dual-shuffle attention network was employed by Chattopadhyay et al. to classify images of breast malignancy. Chattopadhyay et al. used a thick residual dual-shuffle attention network to categorize photos of breast cancer. In Ref[32], a brand-new SAFNet breast cancer detection model is put out, its core model being a pre-trained ResNet-18 paired with the spatial attention mechanism. The regions of interest (RoI) are localized and utilized to steer the attention of the classification network concurrently in the supervised attention approach provided by et al.[33] for the detection of breast cancer histopathology pictures using CNN. The proposed supervised attention mechanism suppresses activations in irrelevant and distracting regions while selectively activating neurons in symptomatically important regions. Breast ultrasound (BUS) picture concurrent segmentation and binary classification using a multi-task learning (SHA-MTL) model is described in [34]. It employs both gentle and firm attention techniques.

Chapter 3: PROPOSED METHODOLOGY

This chapter discusses the methodology used in this research work. It also describes about the dataset and the other related study. It also depicts the comparison and limitation of the traditional method with proposed method. In this chapter it first describes the proposed framework. After that it illustrates about the attention, self attention and limitation of self attention. Then it describes the the limitation of self attention. After that it illustrates the framework of external attention and it's advantages.

3.1 PROPOSED SYSTEM

A block schematic diagram of the complete procedure is displayed in Fig. 1 of the work, which proposes an external attention multi layer perceptron (EAMLP) based transformer technique for the problem of categorizing breast cancer utilizing histological images for binary classification. Two publicly accessible datasets were used in this study. The first is named BreakHis, and the second is called Breast Histology Pictures. The first dataset was utilized for training as well as testing, and the second dataset was used as the test dataset. Initially, the pictures from [62] are labeled as positive and negative for invasive ductal carcinoma (IDC).

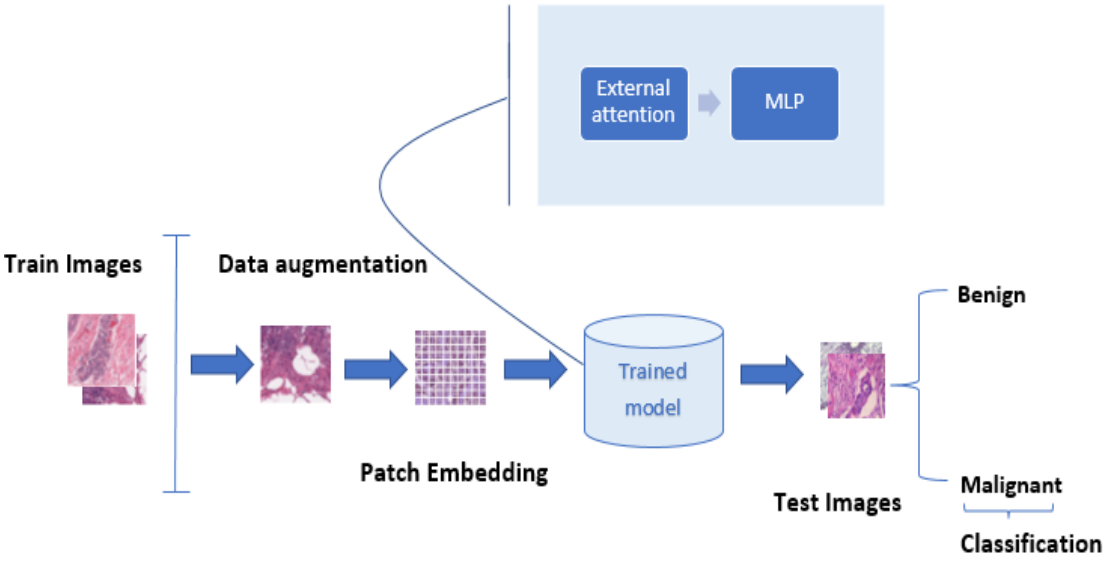


Figure.3.1 Block diagram of proposed model

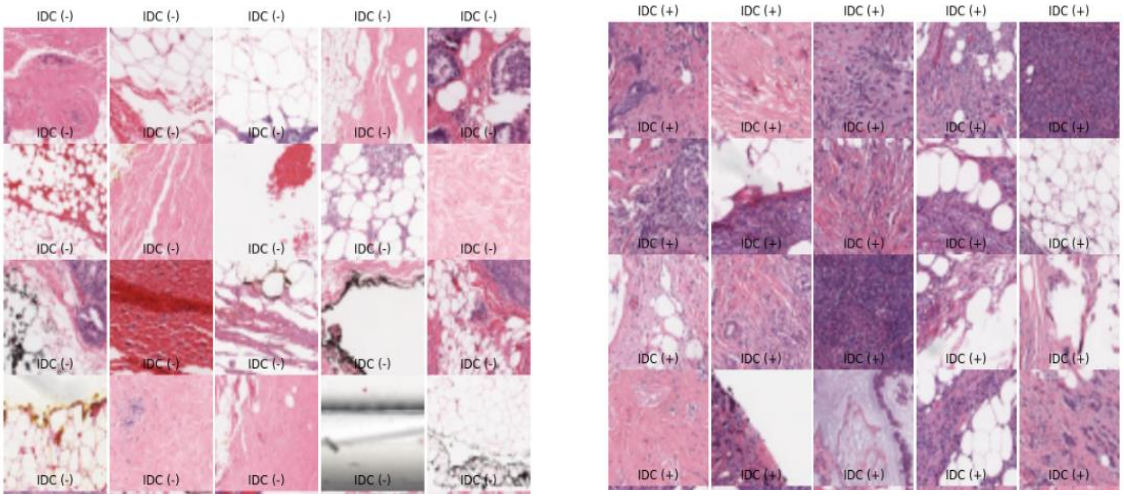
Images which is cancerous is labelled as 1 and images which is non-cancerous is labelled as 0. After labelling pre-processing has been done resizing & data

augmentation technique applied on dataset. For observing the performance based on different image size two image size is used. One is 64x64size & another is 112x112 pixel size. The dataset was then segmented into a train, validation, and test set. After that, the proposed model is trained. Two other models Convolution Neural Network(CNN) & self attention based transformer are applied also for comparative analysis.

3.2 DATA COLLECTION

3.2.1 Breast Histology Images

The Kaggle Breast Cancer Histopathology Images were used to extract the dataset [61]. The collection is made up of 277,524 patches that are each (50×50)pixels in size. Breast cancer specimens were scanned at 40x to provide 162 whole mount slide pictures. 198,738 of the photos had an IDC negative diagnosis, whereas 78,786 had an IDC positive diagnosis. 3,000 IDC positive and 4,500 IDC negative photos make up the subset of 7,500 images used to construct the suggested approach.



IDC + samples IDC - samples

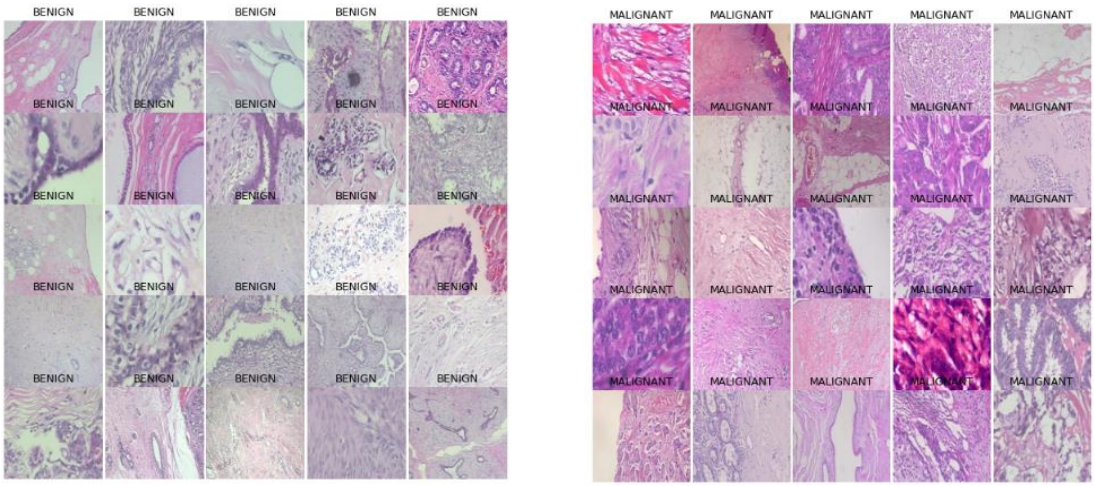
Figure.3.2 Example of image IDC positive and IDC negative Breast Histology Images Datasets

To address the memory error issue, a selection of photographs rather than all of the images was employed[69]. 7500 total photos were used, of which 5250 were used for

training, 1575 for affirmation, and 675 for testing. The breast histopathology image dataset examples are shown in Figure.3.2.

3.2.2 BreakHis Dataset

The BreakHis database contains microbiopsy images of both benign and malignant breast tumors. Pictures were obtained as part of a clinical trial between January and December 2014. Any patients submitted to the P&D Laboratory in Brazil during this time who had a clinical indication of BC were eligible to participate in this trial. The BreakHis database contains microbiopsy images of both benign and malignant breast tumors. Pictures were obtained as part of a clinical trial between January and December 2014. Any patients submitted to the P&D Laboratory in Brazil during this time who had a clinical indication of BC were eligible to participate in this trial.



Benign Malignant

Figure3.3 Example of Benign and Malignant from BreakHis Dataset

There are 2,480 benign and 5,429 malignant specimens with a resolution of 700x460 pixels among 82 people. 1000 photos out of 7909 were utilized just for testing. Examples from the breast histopathology picture dataset are presented in Figure 3.3.

3.3 IMAGE PREPROCESSING

3.3.1 Resized image

The Breast Histology Images' resolution is (50x50). Moreover, the BreakHis dataset's picture resolution is (700 x460). The images in both datasets are altered to

have pixels with the same sizes (64x64) and (112x112). To fix the issue with the poor resolution of the images, the pictures were altered. To assess the model's performance, the images were scaled into two different forms[63]. The dataset is divided into training, validation, and testing datasets before the network is trained.

3.3.2 Data augmentation

To swamped these problems, data augmentation is used. A lot of data is required for CNN to learn its parameters. The training data set can be upgraded via augmentation often [64]. By augmentation, a system's performance can be improved while the danger of overfitting and data imbalance is reduced. There are several ways to improve data. They comprise causal reflection, rotations, and horizontal or vertical translations. The test set was not impacted; only the training set got the data augmentation. It helps machine learning models perform and produce better results. Keras pre-processing layers have been used for data augmentation. The applied augmentation techniques include Random Rotation, RandomZoom , RandomFlip. To create symmetrical data concerning the vertical axis, a horizontal flip is preferably employed.

3.4 SYSTEM ARCHITECTURE

In this section we first know what is attention mechanism, why it is needed. Then we will know about self-attention mechanism used in traditional transformer, the limitation of this self-attention. Then we will know about the proposed model called EAMLP(external attention multi layer perceptron) based transformer. For the comparison purpose the convolution neural network and traditional self-attention based transformer has been used.

3.4.1 Attention Mechanism

A complicated cognitive process in humans is attention. When spotting an item in a visual scene, it's utilized to concentrate on its constituent sections, components, or features [65]. In addition to sensory stimulation , which can cause us to notice features like novelty and unexpectedness, selecting , which might cause us to pay attention, such knowledge, expectancy, and our present objective, can also do so [66]. The resolution of the Breast Histology Pictures is (50x50). Moreover, the image resolution of the BreakHis dataset is (700 x 460). Both dataset's pictures have been modified to use

pixels with the same sizes (64x64) and (112x112). To remedy the problem with the poor resolution of the images, the pictures were altered. To assess the model's performance, the images were transformed into two different forms. The dataset is prorated into training, validation, and testing datasets before the network is trained. The main objective of this function is to strive to emphasize the important information while downplaying the irrelevant details. Working memory is limited in both humans and computers, hence this method is crucial to prevent overtaxing a system's memory. In deep learning, attention may be seen as a vector of significant weights. The relationship between one element and other elements, such as a word in a sentence or a pixel in an image, is predicted using the attention vector[68]. Initially, Bahdanau et al. [69] recommended attention as a means to facilitate neural machine translation's storage of lengthy input sentences. In the past, researchers mostly developed a pattern model, which includes an encoder and a decoder, to address problems with neural machine translation. One context vector is produced by the encoder's eventual concealed state and utilized as the decoder's input. All of the input sequence's contextual information is anticipated to be contained in the context vector. The inability to recall lengthy sentences is thus a significant drawback of this context vector.

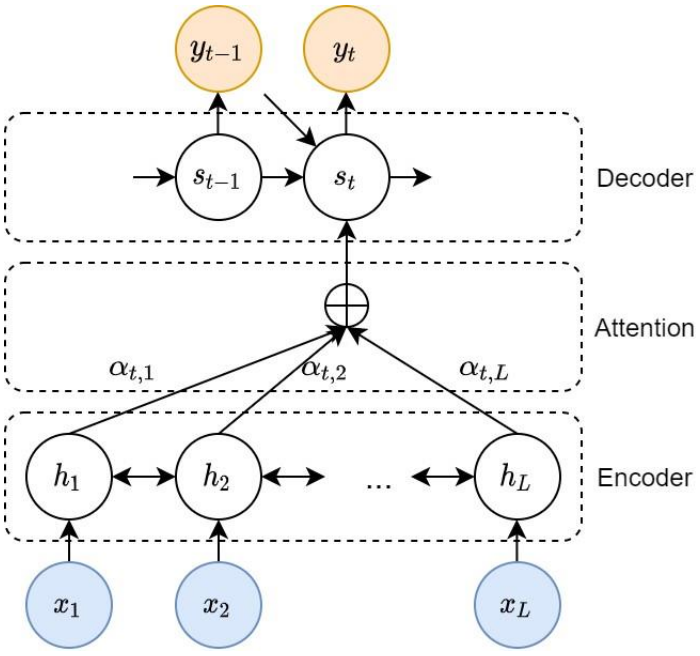


Figure.3.4 Illustration of an encoder-decoder model with attention [69], at time step t; each circle depicts a vector.

The context vector and the entire input sequence are connected via shortcuts produced by the attention mechanism initially presented by Bahdanau et al. [69], and the weights

of these detours are able to be acquired for each output fundamental. The attention mechanism builds the new context vector using the synchronization of the source and target aspects, the decoder hidden states, and the encoder hidden states.

1. **Alignment scores**: With the previously produced decoder , s_{t-1}, and the encoded hidden states ,h_i,the alignment model calculates a score, $e_{t,i}$, that shows how well the input sequence's individual parts match with the current output at the position, t. The alignment model's representational function a (.) may be implemented by a feedforward neural network:

$$e_{t,i} = a(s_{t-1}, h_i) \quad (1)$$

2. **Weights**: By performing a softmax operation on the approximated alignment scores, the weights, $\alpha_{t,i}$ are determined:

$$\alpha_{t,i} = \text{softmax}(e_{t,i}) \quad (2)$$

3. **Context vector**: The decoder gets a unique context vector, c_t. at each time step. The weighted total of all T hidden encoder states is used to compute it.

$$c_t = \sum^{T} \alpha_{t,i} h_i \quad (3)$$

Following are some reasons for using attention for image classification:

- For the purpose of classifying images, attention methods can be employed to guide the model's attention to the most crucial aspects of the picture while disregarding unimportant or noisy details. Attention mechanisms can increase the precision of image classification algorithms by selectively focusing on various areas of the image.
- We want the image model we train to be capable of concentrating on the picture's essential components. One way to do this is through trainable attention processes.
- Traditionally, convolutional neural networks (CNNs) have been used to classify images. These CNNs scan the image using a fixed-sized kernel, which can be limited since it assumes that every part of the image contributes equally to the classification outcome. Contrarily, attention methods enable the model to

dynamically modify the weights assigned to various visual components in accordance with their significance for the classification job.

- The model's categorization judgments may be seen and interpreted using attention processes, which can offer important insights into how well the model is doing and where it might need to be improved.

  For instance, if a car and a backdrop are present in an image, the attention mechanism may be trained to concentrate just on the automobile and ignore the background. By lessening the influence of irrelevant characteristics, this can increase the classification's accuracy.

3.4.2 Self -Attention Mechanism

Transformer is a seq2seq model developed by Vaswani et al. [70] that totally depends on self-attentional mechanisms and uses positional encoding in place of recurrent architecture. An attentional technique called self-attention joins numerous tokens from a single sequence to produce a representation of the same sequence. It may be viewed as a special application of attention when the two sequences to be linked are identical.The queries (Q), keys (K), and values (V) are the three basic parts of the general attention mechanism. Calculating the Query, Key, and Value matrices proceeds first. It is accomplished by multiplying the image's matrix, X, by the weight matrices WQ, WK, and WV which presents in Figure 3.5. The weight matrices in this case are initialized at random.

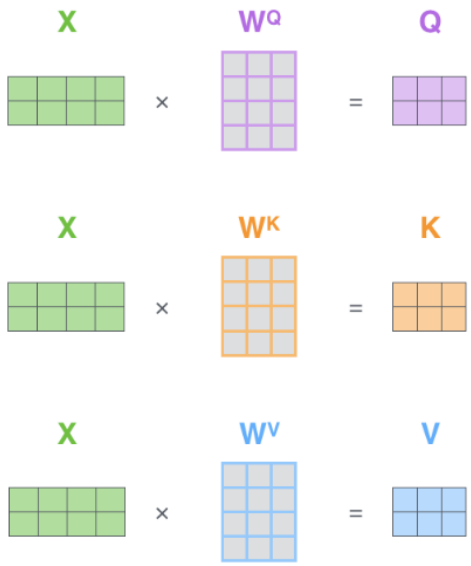


Figure. 3.5 Multiplying image vector with weight matrices[71]

Vaswani et al. [70] created the Transformer seq2seq model, which completely relies on self-attentional processes and employs positional encoding in place of recurrent architecture. Self-attention is an attentional strategy that combines a number of tokens from a single sequence to create a representation of that sequence. If the two sequences that need to be connected are the same, it may be considered a specific application of Attention. The **third steps** are to divide the scores by the square root of the dimension of query and key. This is to allow for more stable gradients, as multiplying values can have exploiding effects. This dimension which is used for dividation has been depended on the dimension of query and key vector. The dimension of both query and key vector is same .

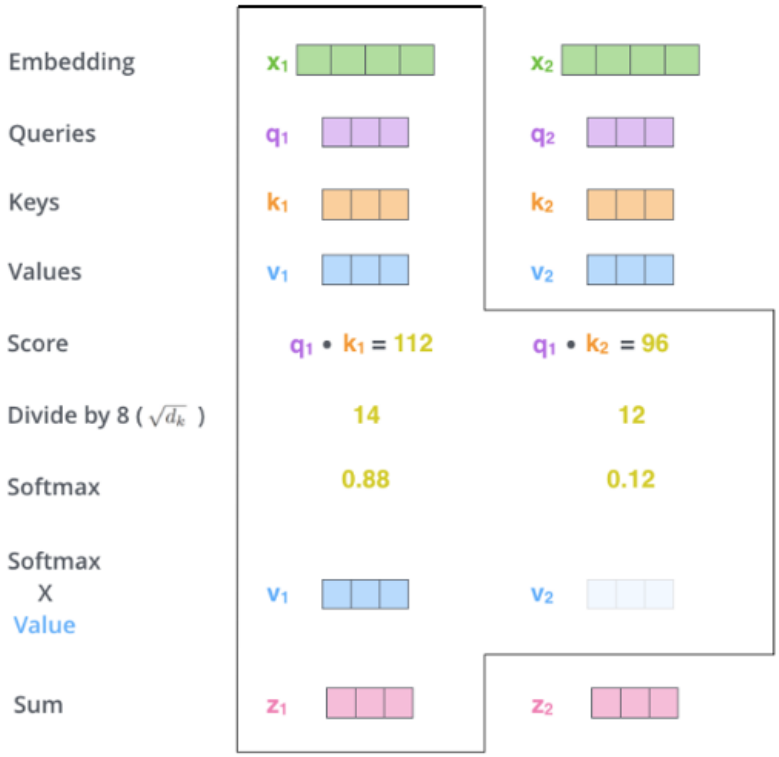


Figure.3.6 Illustration of self attention calculation [71]

After obtaining the softmax of the scaled score, which offers probability values between 0 and 1, the attention weights are then determined in the fourth phase. Higher scores are increased and lower scores are decreased by performing a softmax. With more confidence, the model may decide which picture vector to concentrate on. Step-by-step calculations for the self-attention process are shown in Figure.3.6 .The **fifth step** is to multiply each value vector by the softmax score. The attention weights are then

multiplied by your value vector to get the output vector. The importance of words that the model learns will keep growing the higher the softmax scores. The lower scores will

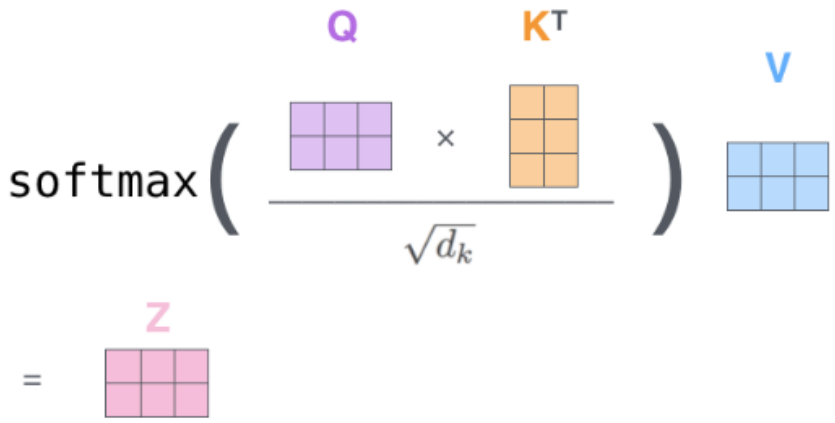


Figure. 3.7 Self-attention calculation in matrix form [71]

obscure the less significant terms. It then sends the output to a linear layer for processing.In the sixth phase, the weighted value vectors are added. In the sixth phase, the weighted value vectors are added. The output for the first picture vector is the self-attention layer as a consequence. The whole step of the self-attention computation is shown in matrix form in Figure 3.7.

3.4.3 Multi -head Attention

A collection of attention processes is combined to form multi-head attention [70]. The heads, or group members, have a similar structure but distinct criteria. Several attention processes can be computed simultaneously. The multi-head attention structure is shown in Figure 3. Each self-aware process is introduced to as a head in the context of self-attention. The output vector of each head is concatenated into a single vector and then sent through the last linear layer. As each head would potentially learn something different, the encoder technique would have a greater representational capacity.

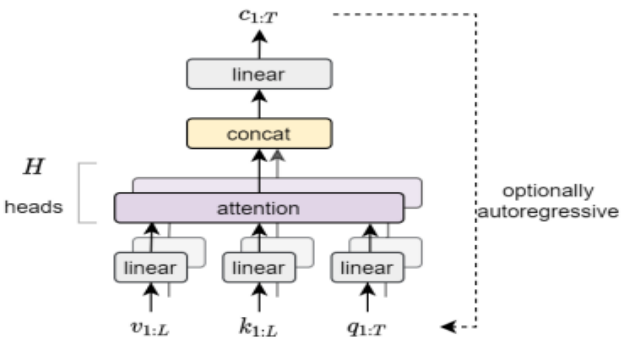


Figure. 3.8 Illustration of multi-head attention [70].

3.4.4Transformer Blocks

A Transformer layer, also known as a Transformer block, is made up of several fundamental modules. Transformer blocks come in two varieties: encoder blocks, which encrypt a sequence, and decoder blocks, which link two sequences. Both Transformer block types are shown in Figure 3.8, along with how they are integrated to create an encoder-decoder model.

Transformer decoders are typically used for language generation tasks such as machine translation, summarization, and text generation. In these tasks, the decoder is responsible for generating the output sequence based on the encoded input sequence and attention mechanism. In contrast, for image classification, the input is a 2D grid of pixel values, not a sequence of tokens. The encoder part of the Transformer is used to process this image and extract relevant features, but the decoder is not used. Instead, the output of the encoder is typically fed into a classifier such as a fully connected layer, which predicts the image class.

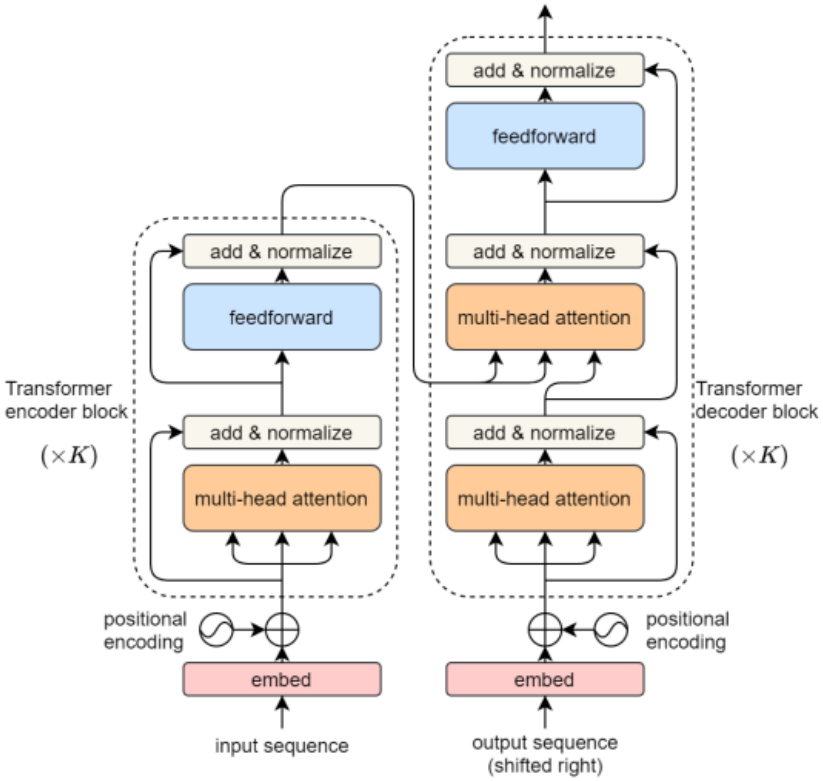


Figure. 3.9 Illustration of Transformer blocks [70].

Therefore, the decoder is not used in Transformer for image classification because the task doesn't require sequence generation and the output is a single class label, not a sequence of tokens. The mechanism of different portion of the encoder for image classification are described below:

- **Patch embedding:**

Patch embedding is a crucial component of the Transformer architecture because it enables the transformer to process image data, which is inherently different from sequential text data that transformers are traditionally used for. In the transformer for image classification, the input image is divided into non-overlapping, fixed-size patches that are then flattened into a collection of vectors. With the use of this patch embedding technique, the Transformer is able to process the picture as a series of embeddings that can be processed by its self-attention mechanism. The transformer can only handle visual data with patch embedding since its self-attention mechanism demands a sequential input. Patch embedding therefore serves as a bridge between image data and the transformer architecture, allowing the transformer to process images as effectively as it does sequential data. In this work, patch size 2×2 has been used for patch embedding.

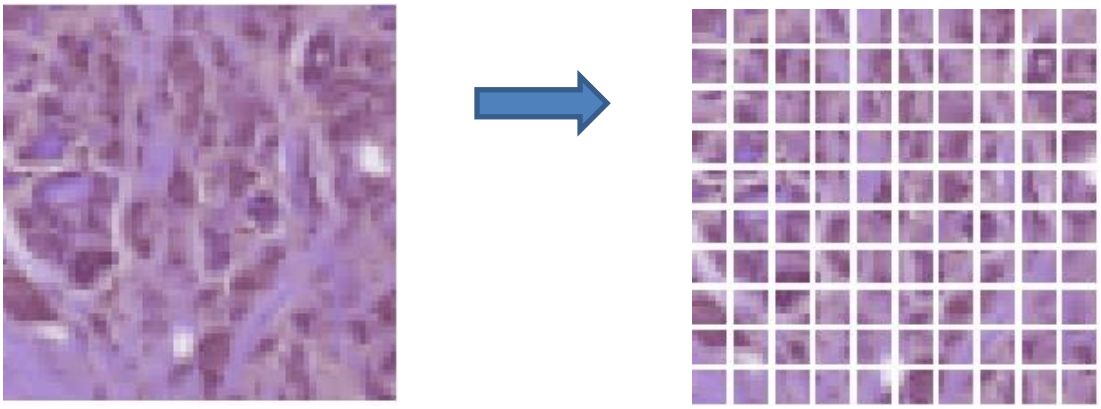


Figure. 3.10 Patch embedding of breast image

- **Positional Embedding:**

The embeddings' next stage is to receive positional information. As the transformer encoder lacks recurrence in contrast to recurrent neural networks, some positional information must be provided in the input embeddings. Positional encoding is used to achieve this. There is a cos vector for each odd location in the input vector. The sin

function is used to each even index to produce a vector. After that, add the necessary vectors to the input embeddings. This successfully informs the network of each vector's

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \quad (4)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \quad (5)$$

location. Sine and cosine functions have been merged because they both have immediately evident linear features.

- **Encoder Layer:**

The encoder layers' job is to convert each iteration process into an abstract supervised learning approach that comprises the entire sequence's learned data. There are two submodules, several heads of attention, and an entirely integrated network. Residual connections are found close to each of the two sublayers after a layer normalization[79].

- **Multi Head Attention:**

Multi-headed attention in the encoder uses a specific attention mechanism called self-attention. For self-attention, the models are able to connect every word in the input to every other word. Mechanisms for self-attention and multi-head attention have already been covered in the preceding section.

- **The Residual Connections, Layer Normalization, and Feed Forward Network:**

The fundamental positional input embedding is expanded with the multi-headed attention output vector. This is a residual connection, as the name implies. The output of the residual connection is subjected to layer normalization. A pointwise feed-forward network is proposed to process the normalized residual output further. Each of the few linear layers that make up the pointwise feed-forward network is separated by a ReLU activation. The output is further normalized before being added to the pointwise feed-forward network's input. By enabling gradients to move directly across the networks, the remaining connections aid in network training. The network is stabilized using layer normalizations, which significantly reduces the amount of training time required. A richer representation may come from projecting the attention outputs onto the pointwise feedforward layer[72].

The encoder layer is finished with that. The goal of each of these procedures is to encrypt the input for a continuous representation with attention information.

3.4.4 Limitation of self-attention mechanism

For visual tasks, deep feature representation has grown more reliant on attentional mechanisms, notably self-attention. Self-attention uses pair-wise affinities across all locations to estimate a weighted sum of characteristics in order to represent the long-range dependency inside a single sample. Yet, self-attention is quadratic complicated and excludes the possibilities of sample correlation.

Let assume, X is set of some set of inputs where $X \in \mathbb{R}^{N\times d}$ and $W_q \in \mathbb{R}^{d\times d}$, $W_k \in \mathbb{R}^{d\times d}$, $W_v \in \mathbb{R}^{d\times d}$, are the matrices with learnable parameters . Multiplication between them computes the query vector first ($Q = XW_q$), key ($K = XW_k$), and value ($V = XW_v$) matrices respectively. Here $d_q = d_k$, is assumed . The Q and K is then $N \times d_k$,in their sizes, while the V is $N \times d_v$, in it's sizes. Where, N is the number of patch & d is the embedding size. The premise of the softmax dot-product self-attention operation is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{6}$$

Here, $d=d_k=d_q$. Let Y be a standardization function of some type. The following is a definition of self-attention that is more comprehensive:

Consider the scenario where N is significantly larger to d_k ,d_v. Following is the calculation of computational complexity of self attention mechanism:

$$\text{Attention}(Q,K,V) = Y(S) \cdot V. \tag{7}$$

1) calculation of $Y=QK^T/\text{sqrt}(d)$ takes $O(N^2d)$.

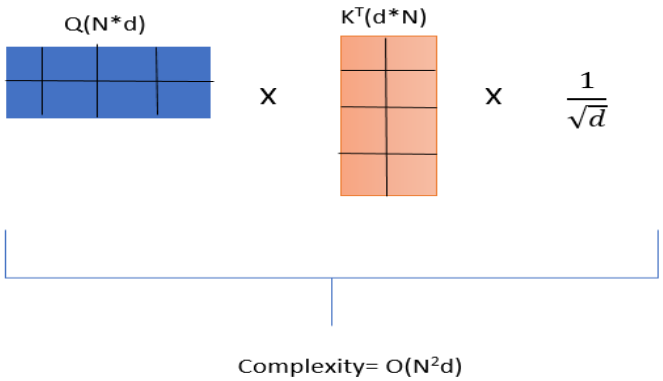


Figure.3.11 Illustration of computational complexity of self attention mechanism

2) It takes $O(N^2)$ time to exponentiate and calculate the row sum of N,

3) it takes O to divide each element of N by the appropriate row sum (N^2), and

4) Softmax(QK^T) and V are multiplied in $O(N^2dv)$ time. As a result, the computational difficulty of this crude method for computing self-attention scales quadratically in N[80]. The authors claim that since the majority of pixels are closely related to relatively few other pixels, an N-to-N attention matrix may be unnecessary. Self-attention may be limited in its flexibility and capacity since it only takes into account correlations among items within a single data sample and overlooks possible relationships among elements across other samples.

3.4.4 External Attention Multi Layer Perceptron(EAMLP) Based Transformer

In this work, a brand-new attention mechanism called "external attention" is presented for the classification of breast cancer images. It occupies the place of self-attention in existing conventional designs. With just two cascaded linear layers and two normalizing layers, it is easily implementable. Its foundation is made up of two external, minute, teachable shared retention units. The correlations between all data samples are implicitly taken into account by external attention, which is linearly sophisticated. In the interest of developing an all-MLP architecture for breast image classification called external attention MLP (EAMLP), we also implement the multi-head mechanism into external attention.

3.4.4.1 External Attention

External attention is a type of attention which consists of two retention unit. The first step in computing self-attention is to generate an attention map by figuring out the affinities between the self query and self key vectors. The subsequent phase involves developing a new feature map by incorporating this attention map into the scaling factor of the self value vectors. Several processes govern external attention. To boost the network's capacity, it employs two separate retention units, Rk and Rv, as that of the key and value for external attention. R is a characteristic that can be learnt in this case and acts as a storage for the whole training dataset regardless of the input. The computation of external attention can be written as:

$$A = \text{Norm}(XR_k^T) \tag{8}$$

$$X_{out} = AR_v \tag{9}$$

Algorithm 1 external attention.

Function: external_attention

Inputs:

- x: tensor of shape (batch_size, num_patches, channel)
- dim: integer, the dimensionality of the output space
- num_heads: integer, the number of attention heads
- dim_coefficient: integer, the coefficient to increase the number of attention heads
- attention_dropout: float, dropout rate for attention weights
- projection_dropout: float, dropout rate for projection output

Outputs:

- x: tensor of shape (num_patch, channel)
- update num_heads to (num_heads * dim_coefficient).
- x=query_linear(dim * dim_coefficient) x.
- x=Reshape x (-1, num_patch, num_heads, dim * dim_coefficient // num_heads).
- x=x . permute (0, 2, 1, 3).
- attn = query_linear (dim // dim_coefficient) x. → R_k
- attn= softmax (attn ,axis=1) .
- attn=attn/(1e-9 + the sum of attn along the last dimension)
- attn = Dropout(attention_dropout)(attn)
- x = query_linear (dim * dim_coefficient // num_heads)(attn) → R_v
- x =x. permute (0, 2, 1, 3).
- x =reshape(x, [-1, num_patch, dim * dim_coefficient])
- x = query_linear (dim)(x)
- x = Dropout(projection_dropout)(x)

As seen in Algorithm 1 below, it is simple to accomplish by utilizing only two linear layers and two normalizing layers. Here, Rk stands for the retention unit that uses a key vector instead of a value vector, and Rv for a value vector instead of a key vector. In contrast to self attention, two retention units in external attention are implemented without employing bias. With the help of two linear layers and two normalizing layers, this feature map has incorporated the retention unit. By lowering complexity, it can therefore take the role of the self-attention in the transformer.

3.4.4.2 Computational Complexity of External Attention

Let assume X is the set of inputs where $X \in \mathbb{R}^{N \times d}$ & N is the number of patch & d is the embedding size. The external-attention action for the softmax dot-product is described as:

$$A = \text{Norm}(XR_k^T)$$

$$X_{out} = A\, R_v$$

where retention unit $R \in \mathbb{R}^{S \times d}$. For external attention, the computational complexity is thus:

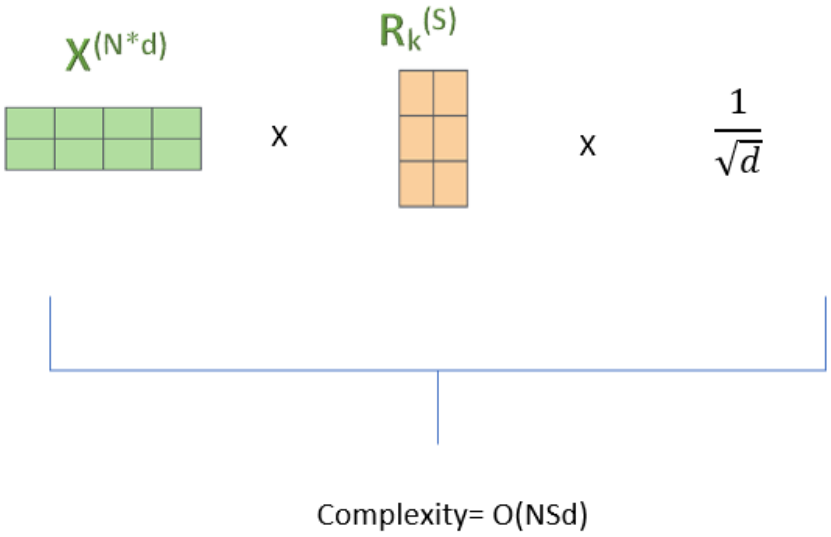


Figure.3.12 Illustration of computational complexity of external attention mechanism

The recommended method is linear in the pixel density and has an external attention computational complexity of O(NSd) since d and S are hyper-parameters. As d and S are hyper-parameters, the recommended method is linear in the quantity of pixels. This is equivalent to a drop patch action since a lot of the data in a patch in an image is irrelevant and redundant.External attention is substantially more efficient than self-attention because of the reduced computational cost.

3.4.4.3 Double Normalization

Normalization is utilized to solve the problem of disappearing gradients that might occur because of the deep structure of the model in the setting of self-attention in transformers. Likewise in the case of external attention first normalization has been

done by using softmax function. It normalizes the product of input feature map and the transpose of retention unit R_k using softmax function.

$$A = \text{Norm}(XR_k^T) \longrightarrow \text{First Normalization}$$

The attention map is based on the size of the input features, as opposed to cosine similarity. To go around this difficulty, double standardization has been used[80] on the attention map.

attn=attn/(1e-9 + the sum of attn along the last dimension)⟶ Double Normalization.

The first term in the double normalization equation, which is subsequently utilised to compute the attention map using the softmax function, is the scaled dot-product of the normalized retention(Rk) vectors and the normalized feature map. The second term is the normalized attention map. The second normalizing of the attention map is meant to assist the model in focusing on pertinent data in the input sequence, while the combined normalization of the feature map and retention(R_k) vectors is meant to increase the model's stability and generalizability.

3.4.4.4 Multi Layer Perceptron

A completely linked multi-layer neural network is referred to as a "Multilayer Perceptron" (MLP). A sort of feedforward artificial neural network called a multilayer perceptron produces outputs from a collection of inputs. There are several levels of input nodes in the directed graph that connects a multilayer perceptron's input and output layers. A deep learning method called backpropagation is utilized to train the Multilayer Perceptron network. Many studies demonstrate that multilayer perceptrons alone are unable to fully account for the success of transformers (MLPs)

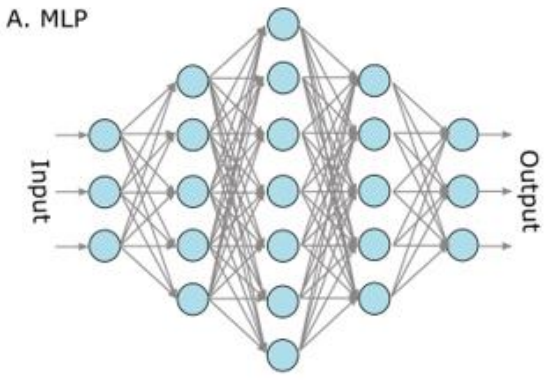


Figure.3.13 Illustration of MLP[76]

The addition of a multi-layer perceptron (MLP) or dense layer in self-attention-based transformers serves two main purposes:

Non-linear transformation: Self-attention only captures linear relationships between tokens in a sequence. However, many natural language processing (NLP) tasks require non-linear relationships to be captured. By adding an MLP or dense layer after the self-attention layer, the model can learn non-linear transformations of the self-attention output, which can be useful for capturing more complex relationships between tokens[73].

Feed-forward network: The self-attention layer's output can be processed by a feed-forward network that is part of the MLP or dense layer. This enables the model to make a final prediction after doing some further processing on the resultant representations in addition to capturing the connections between the tokens in a sequence. This extra processing can help the model perform more accurately on a variety of picture categorization tasks[74].

Overall, self-attention-based transformers can become more potent and successful for a variety of picture classification applications by including an MLP or dense layer after the self-attention layer.

Chapter 4: EXPERIMENTS AND RESULT ANALYSIS

The outcomes of trials using CNN and two attention-based transformers for the classification of breast cancer using two datasets are presented in this chapter. The suggested model is tested using several MLP layers with various pixel sizes. The binary classification result will be described in the following sections, along with a comparison to other research' results.

4.1 PERFORMANCE OF PROPOSED MODEL

The classification result has been depicted based on different parameter. In this work one dataset has been used for train and test the model, another is used to test the model. Moreover two different pixel size 64×64 & 112×112 has been used in this research work for observing the performance of the model. Here the dense layer is hyperparameter. This performance of the proposed model has been observed based on different MLP. For breast histopathology images the accuracy, specificity and sensitivity has been objected in Table 4.1 & Table 4.3 for different pixel size. Similarly For breakHis datasets the accuracy, specificity and sensitivity has been objected in Table 4.2 & Table 4.4 for different pixel size. The values of true false positives and negatives are used to determine all of the parameters.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$\text{Sensitivity}=\frac{TP}{TP+FN} \tag{11}$$

$$\text{Specificity}=\frac{TN}{TN+FP} \tag{12}$$

- TN= True Negative.
- TP= True Positive.
- FN= False Negative.
- FP= False Positive.

Table 4.1 shows the performance of breast histopathology images for 64×64 pixel size based on different number of dense layer. The performance has been observed based on accuracy, sensitivity, specificity. Figure 4.1 depicts the comparison chart based on number of dense layers for breast histopathology images where pixel size is 64 x 64. From Table 4.1 it has been observed that the higher accuracy has been achieved 94.22% where number of dense layer is 4.

Table 4.1 Performance Measures Of Proposed Model Using Breast Histopathology Images For 64×64 Pixel Size

Model	Dataset	Pixel size	No of dense layer	Accuracy(%)	Sensitivity(%)	Specificity(%)
External Attention based Transformer	Breast Histopathology Images	64x64	2	92.59	99.73	84.14
		64x64	3	91.56	98.14	83.33
		64x64	4	94.22	97.74	88..29
		64x64	5	92.59	97.88	85.62
		64x64	6	94.07	92.76	96.64
		64x64	7	81.33	98.97	60.46

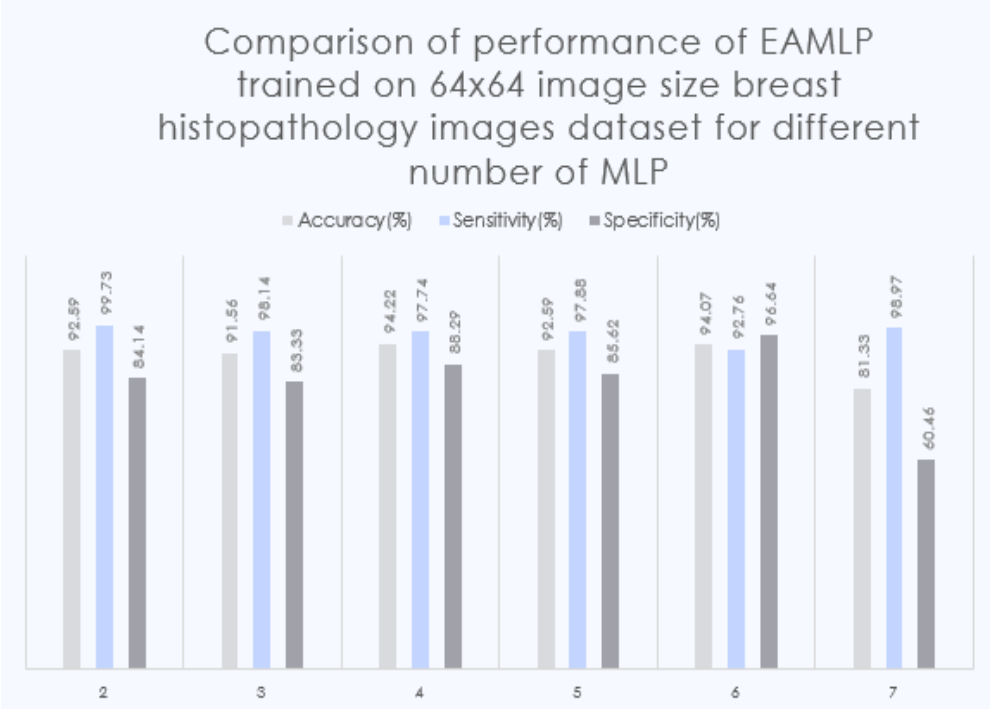


Figure.4.1 Performance of EAMLP trained on 64x64 image size Breast Histology Images for different number of MLP

Table 4.2 shows the performance of breakHis datasets for 64×64 pixel size based on different number of dense layer. The performance has been observed based on accuracy, sensitivity, specificity. Figure 4.2 depicts the comparison chart based on number of dense layers for breakHis datasets where pixel size is 64 x 64. From Table 4.1 it has been observed that the higher accuracy has been achieved 94.47% where number of dense layer is 6.

Table 4.2 Performance Measures Of Proposed Model Using BreakHis Dataset For 64×64 Pixel Size

Model	Dataset	Pixel size	No of dense layer	Accuracy(%)	Sensitivity(%)	Specificity(%)
External Attention based Transformer		64x64	2	92.73	99.15	85.28
		64x64	3	91.13	97.78	82.35
	BreakHis	64x64	4	93.59	97.03	88..88
		64x64	5	92.59	97.42	86.29
		64x64	6	94.47	93.65	95.94
		64x64	7	81.99	98.83	69.04

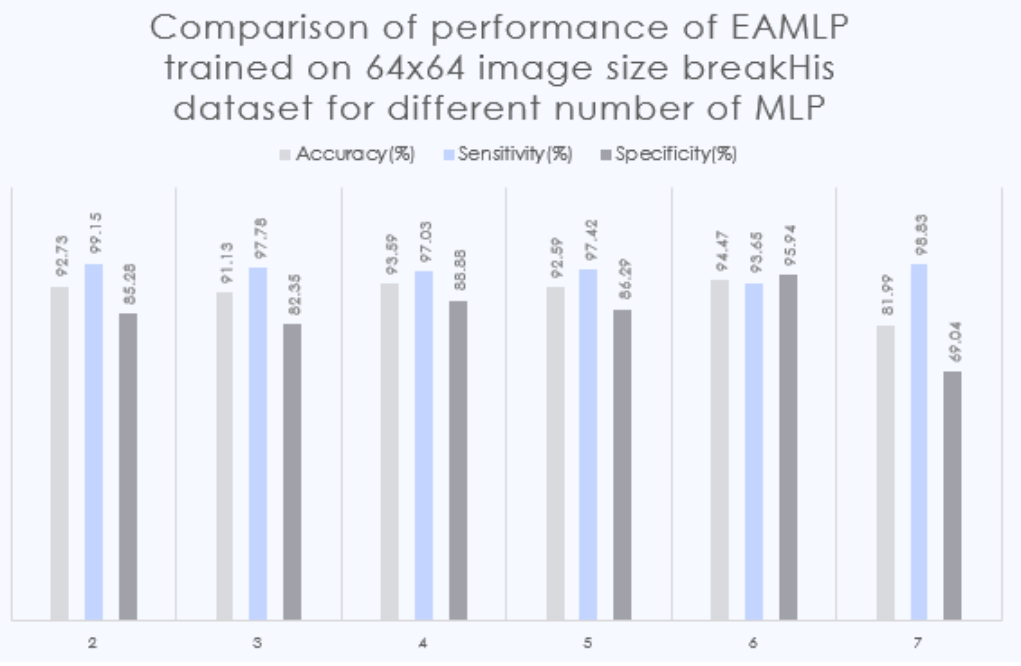


Figure.4.2 Performance of EAMLP trained on 64x64 image size BreakHis datasets for different number of MLP

Table 4.3 shows the performance of breast histopathology images for 112×112 pixel size based on different number of dense layer. The performance has been observed based on accuracy, sensitivity, specificity. Figure 4.3 depicts the comparison chart based on number of dense layers for breast histopathology images where pixel size is 64 x 64. From Table 4.3 it has been observed that the higher accuracy has been achieved 94.52% where number of dense layer is 6.

Table 4.3 Performance Measures Of Proposed Model Using Breast Histopathology Images For 112×112 Pixel Size

Model	Dataset	Pixel size	No of dense layer	Accuracy(%)	Sensitivity(%)	Specificity(%)
External Attention based Transformer	Breast Histopathology Images	112x112	2	93.15	96.23	91.12
		112x112	3	93.77	95.82	90.56
		112x112	4	89.18	98.87	79.02
		112x112	5	94.22	94.11	94.26
		112x112	6	94.52	97.74	89.93
		112x112	7	88.44	99.12	77.61

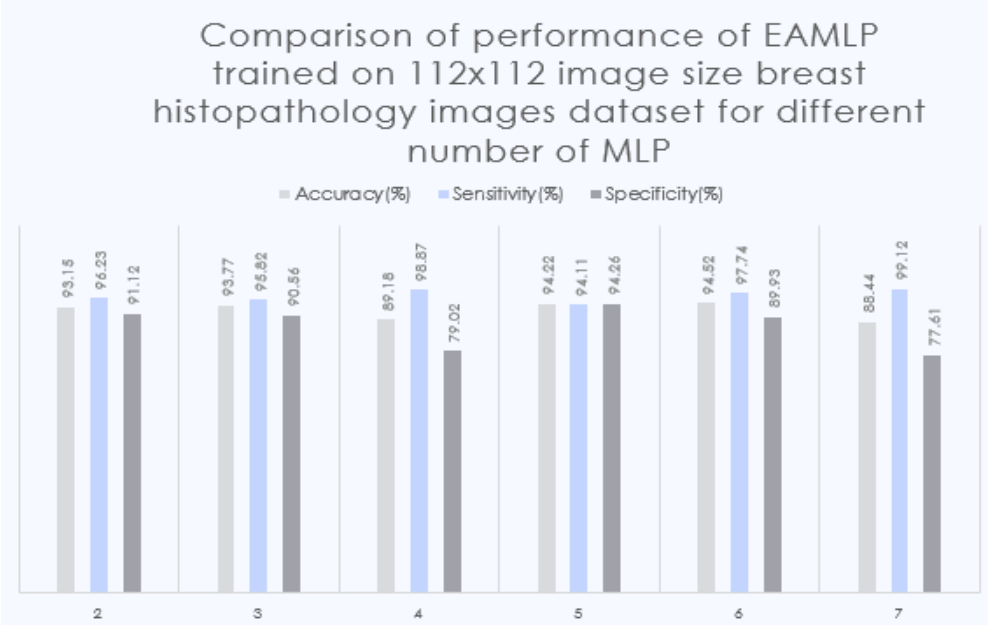


Figure.4.3 Performance of EAMLP trained on 112x112 image size Breast Histology Images for different number of MLP

Table 4.4 shows the performance of breakHis datasets for 112×112 pixel size based on different number of dense layer. The performance has been observed based on accuracy, sensitivity, specificity. Figure 4.4 depicts the comparison chart based on number of dense layers for breast histopathology images where pixel size is 64 x 64. From Table 4.4 it has been observed that the higher accuracy has been achieved 95.73% where number of dense layer is 5.

Table 4.4 Performance Measures Of Proposed Model Using BreakHis Dataset For 112×112 Pixel Size

Model	Dataset	Pixel size	No of dense layer	Accuracy(%)	Sensitivity(%)	Specificity(%)
External Attention based Transformer	BreakHis	112x112	2	93.77	96.19	90.29
		112x112	3	92.93	95.18	89.55
		112x112	4	89.73	98.72	80.28
		112x112	5	95.73	96.11	95.16
		112x112	6	94.53	97.93	89.74
		112x112	7	87.53	98.79	76

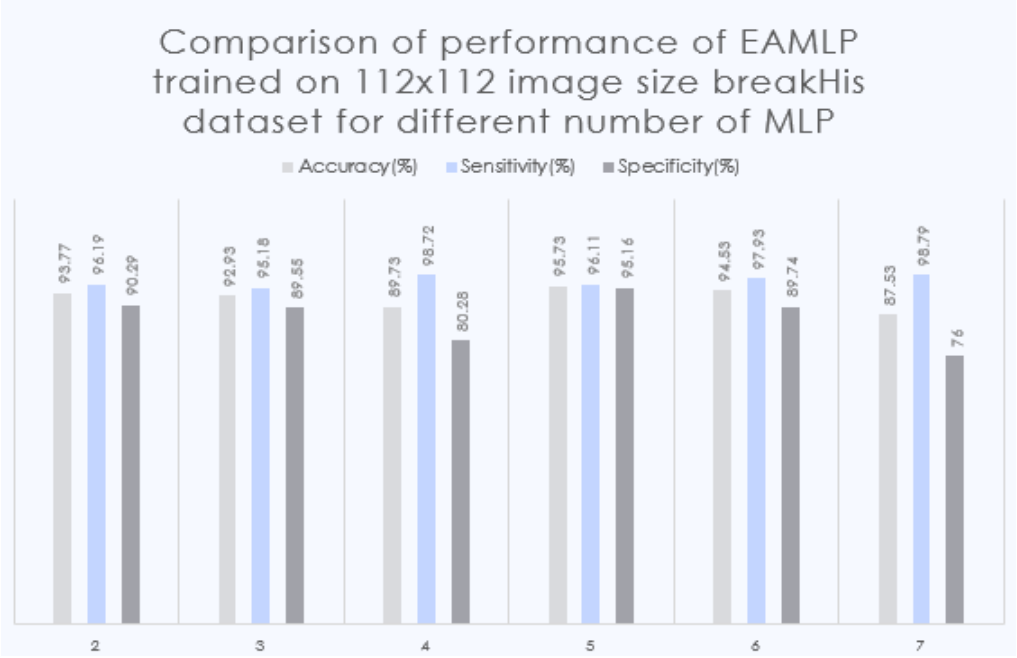


Figure.4.4 Performance of EAMLP trained on 112x112 image size BreakHist datasets for different number of MLP

Table 4.5 Performance Measures Of Proposed Model With Traditional Approach

Model	Accuracy(%)
self -attention	90.96
External Attention for Breast Histology Images On 64x64 pixel size	94.22
External Attention for BreakHis Dataset On 64x64 pixel size	94.47
External Attention for Breast Histology Images On 112x112 pixel size	94.52
External Attention for BreakHis Dataset Images On 112x112 pixel size	95.73

Finally, for each data set with its corresponding picture size, the best accuracy value accountable for that specific MLP has been highlighted. The classic self-attention based transformer shown in Table 4.5 was then contrasted with the accuracy value that had been emphasized. According to these tables, external attention with MLP 5 has the higher accuracy (95.73%) in comparison to other tables & traditional approach.

Table 4.6 Comparison Of The Proposed Method With Existing Work

Reference	Method	Accuracy
[32]	SAF_NET	94.10%
[33]	Guided soft attention	93%
[34]	SHA-MTL	94.12%
[35]	Spatial Attention Guided	93%
[75]	Attention gate	87%
[76]	Region Guided Attention	88%
Proposed model	EAMLP	95.73%

4.2 COMPARISON OF PROPOSED MODEL WITH EXISTING WORK

The proposed strategy is contrasted with the current approaches to confirm the effectiveness of the classification. Based on the feature extraction methodology, classification method, and performance parameter, Table 4.6 compares the proposed study with a few other research efforts that have already been published. This table demonstrates how much the strategy suggested in this paper improves classification accuracy. In comparison to prior work shown in the table, the suggested strategy is therefore capable of greatly improving classification accuracy.

CHAPTER 5:CONCLUSION

This chapter includes the summary about the thesis work. It depicts the short description of the whole work and analyse the methodology. It also discuss about the result and illustrates the best result. Moreover, it discusses the challenges of the thesis work. Then it illustrates some future work for solving this limitation.

5.1 CONCLUSION

This study demonstrated an external attention based transformer method for breast cancer image classification . The approached methodology used two linear, small retention units which has been initialized and utilized without using bias. Additionally Multi layer perceptron(MLP) has been employed to extract useful features. Different number of MLP has been employed for different pixel size. For this work, two publicly available datasets has been used which is Breast Histology Images and BreakHis datasets. Here breakHis datasets has been used as test images. This model's performace has been observed by using evaluation matrices accuracy,sensitivity,specificity,f1 score. By observing the performance for different MLP with different pixel size it has been observed that external attention with MLP having 5 layer of 112x112 pixel size has given the better accuracy 95.73% compare with the other model and state of the art. As a part of future work ,the goal would be to implement the model for increasing the performance of the attention based transformer by reducing complexity as much can possible also increasing the pixel size.

5.2 LIMITATION

In this work, the model could not be trained for higher pixel size for lack of storage. Moreover, for higher pixel size more epochs can't be initialized due to storage problem. Additionally, this model performs only for two datasets which are imbalance. For imbalance datasets class wise performance are varied.

5.3 FUTURE STUDY

This model could be trained for higher pixel for further improvement. Moreover, more datasets can be used for observing the performance of the model. More customization can be done for improve the performance and reduce the complexity.

Bibliography

1. https://www.cancer.gov/about-cancer/understanding/what-is-cancer
2. Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71.3 (2021): 209-249.
3. Malon, Christopher, et al. "Identifying histological elements with convolutional neural networks." *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. 2008.
4. "Invasive Ductal Carcinoma: Diagnosis, Treatment, and More." Breastcancer.org, 9 Mar. 2019, www.breastcancer.org/symptoms/types/idc.
5. Elston, C. and Ellis, I., "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up," Histopath. 19(5), 403– 410 (1991).
6. Weston, J., Bengio, S., and Usunier, N., "Large scale image annotation: Learning to rank with joint wordimage embeddings," Machine Learning 81, 21–35 (Oct. 2010).
7. Seide, F., Li, G., and Yu, D., "Conversational speech transcription using context-dependent deep neural networks," in [Proc Interspeech], 437–440 (2011).
8. Glorot, X., Bordes, A., and Bengio, Y., "Domain adaptation for largescale sentiment classification: A deep learning approach," in [Proceedings of the Twenty-eight International Conference on Machine Learning (ICML'11)], 27, 97–110 (June 2011).
9. Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P., "Modeling temporal dependencies in highdimensional sequences: Application to polyphonic music generation and transcription," in [Proceedings of the Twenty-nine International Conference on Machine Learning (ICML'12)], ACM (2012).
10. Ciresan, D. C., Meier, U., and Schmidhuber, J., "Multi-column deep neural networks for image classification," in [Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)], CVPR '12, 3642–3649, IEEE Computer Society, Washington, DC, USA (2012).
11. Mugahed A Al-antari, Mohammed A Al-masni, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. International journal of medical informatics, 117:44–54, 2018.

12. Kassani, Sara Hosseinzadeh, et al. "Breast cancer diagnosis with transfer learning and global pooling." *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2019.
13. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton." Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012. https://doi.org/10.1145/3065386.
14. Simonyan, Karen, and Andrew Zisserman." Very deep convolutional networks for largescale image recognition." arXiv preprint arXiv:1409.1556 (2014).
15. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun." Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016. https://doi.org/10.1109/cvpr.2016.90
16. Szegedy, Christian, Vincent Vanhoucke, Serge Ioffe, Jon Shlens, and Zbigniew Wojna." Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016. https://doi.org/10.1109/cvpr.2016.308
17. Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger." Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017. https://doi.org/10.1109/cvpr. 2017.243
18. Jiang, Y.; Chen, L.; Zhang, H.; Xiao, X. Breast Cancer Histopathological Image Classification Using Convolutional Neural Networks with Small SE-Resnet Module. PLoS ONE 2019, 14, e0214587. [CrossRef]
19. Yang, H.; Kim, J.-Y.; Kim, H.; Adhikari, S.P. Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images. IEEE Trans. Med. Imaging 2020, 39, 1306–1315. [CrossRef] [PubMed]
20. Xu, B.; Liu, J.; Hou, X.; Liu, B.; Garibaldi, J.; Ellis, I.O.; Green, A.; Shen, L.; Qiu, G. Attention by Selection: A Deep Selective Attention Approach to Breast Cancer Classification. IEEE Trans. Med. Imaging 2020, 39, 1930–1941.
21. G. Li, C. Li, G. Wu, D. Ji, and H. Zhang, "Multi-view attention-guided multiple instance detection network for interpretable breast cancer histopathological image diagnosis," IEEE Access, vol. 9, pp. 79671–79684, 2021
22. Xu, Bolei, et al. "Look, investigate, and classify: a deep hybrid attention method for breast cancer classification." *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019.
23. Xu, Bolei, et al. "Attention by selection: A deep selective attention approach to breast cancer classification." *IEEE transactions on medical imaging* 39.6 (2019): 1930-1941.

24. Chen, Gongping, et al. "AAU-net: An Adaptive Attention U-net for Breast Lesions Segmentation in Ultrasound Images." *IEEE Transactions on Medical Imaging* (2022).
25. Biswas, Ananna, Zabir Al Nazi, and Tasnim Azad Abir. "Invasive Ductal Carcinoma Detection by A Gated Recurrent Unit Network with Self Attention." *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2019.q TY23
26. Ukwuoma, Chiagoziem C., et al. "Multi-Classification of Breast Cancer Lesions in Histopathological Images Using DEEP_Pachi: Multiple Self-Attention Head." *Diagnostics* 12.5 (2022): 1152.
27. Xu, Meng, Kuan Huang, and Xiaojun Qi. "Multi-task learning with context-oriented self-attention for breast ultrasound image classification and segmentation." *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022.
28. Al-Hejri, Aymen M., et al. "ETECADx: Ensemble Self-Attention Transformer Encoder for Breast Cancer Diagnosis Using Full-Field Digital X-ray Breast Images." *Diagnostics* 13.1 (2022): 89.
29. Chen, Dehua, and Orlando Mayugi. "Breast tumor diagnosis via phrase level self-attention mechanism." *International Conference on Biomedical and Intelligent Systems (IC-BIS 2022)*. Vol. 12458. SPIE, 2022.
30. Y. Zou, J. Zhang, S. Huang, and B. Liu, “Breast cancer histopathological image classification using attention high order deep network,” International Journal of Imaging Systems and Technology, vol. 32, no. 1, pp. 266–279, 2022.
31. S. Chattopadhyay, A. Dey, P. K. Singh, and R. Sarkar, “DRDA-Net: dense residual dual-shuffle attention network for breast cancer classification using histopathological images,” Computers in Biology and Medicine, vol. 145, Article ID 105437, 2022.
32. Lu, Si-Yuan, Shui-Hua Wang, and Yu-Dong Zhang. "SAFNet: A deep spatial attention network with classifier fusion for breast cancer detection." *Computers in Biology and Medicine* 148 (2022): 105812.
33. Yang, Heechan, et al. "Guided soft attention network for classification of breast cancer histopathology images." *IEEE transactions on medical imaging* 39.5 (2019): 1306-1315.
34. Zhang, Guisheng, et al. "SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification." *International Journal of Computer Assisted Radiology and Surgery* 16 (2021): 1719-1725.

35. Duanmu, Hongyi, et al. "A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images." *Bioinformatics* 38.19 (2022): 4605-4612.
36. J. E. Joy, E. E. Penhoet, and D. B. Petitti, Eds., Saving women's lives: strategies for improving breast cancer detection and diagnosis. Washington, D.C.: National Academies Press, 2005.
37. B. Stenkvist et al., "Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations," Cancer Research, vol. 38, no. 12, pp. 4688–4697, 1978.
38. Breastcancer.org. (2012) Biopsy. [Online]. Available: http://www. breastcancer.org/symptoms/testing/types/biopsy
39. R. Rubin, D. S. Strayer, and E. Rubin, Eds., Rubin's Pathology Clinicopathologic Foundations of Medicine, 6th ed. Philadelphia: Lippincott Williams & Wilkins, 2012. [6] S. R. Lakhani et al., WHO classification of tumours of the breast, 4th ed. Lyon: WHO Press, 2012.
40. S. R. Lakhani et al., WHO classification of tumours of the breast, 4th ed. Lyon: WHO Press, 2012.
41. M. N. Gurcan et al., "Histopathological image analysis: A review," IEEE Reviews in Biomedical Engineering, vol. 2, pp. 147–171, 2009.
42. C. Desir et al., "Classification of endomicroscopic images of the lung based on random subwindows and extra-trees," IEEE Transaction on Biomedical Engineering, vol. 59, no. 9, pp. 2677–2683, 2012.
43. Ting, Fung Fung, Yen Jun Tan, and Kok Swee Sim. "Convolutional neural network improvement for breast cancer classification." Expert Systems with Applications 120 (2019): 103-115.
44. Han, Zhongyi, et al. "Breast cancer multi- classification from histopathological images with structured deep learning model." Scientific reports 7.1 (2017): 1- 10.
45. Kate, Vandana, and Pragya Shukla. "Breast Cancer Image Multi-Classification Using Random Patch Aggregation and Depth-Wise Convolution based DeepNet Model." (2021): 83-100.
46. Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 71.3 (2021): 209-249.
47. Kassani, Sara Hosseinzadeh, et al. "Breast cancer diagnosis with transfer learning and global pooling." 2019 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2019.

48. Malon, Christopher, et al. "Identifying histological elements with convolutional neural networks." Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology. 2008.
49. Cireşan, Dan C., et al. "Mitosis detection in breast cancer histology images with deep neural networks." International conference on medical image computing and computer-assisted intervention. Springer, Berlin, Heidelberg, 2013.
50. Le, Q., Han, J., Gray, J., Spellman, P., Borowsky, A., and Parvin, B., “Learning invariant features of tumor signatures,” in [Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on], 302–305 (May 2012).
51. Malon, C., Miller, M., Burger, H. C., Cosatto, E., and Graf, H. P., “Identifying histological elements with convolutional neural networks,” in [Proceedings of the 5th International Conference on Soft Computing As Transdisciplinary Science and Technology], CSTST ’08, 450–456, ACM, New York, NY, USA (2008).
52. Malon, C. and Cosatto, E., “Classification of mitotic figures with convolutional neural networks and seeded blob features,” Journal of Pathology Informatics 4(1), 9 (2013).
53. Ciresan, D., Giusti, A., Gambardella, L., and Schmidhuber, J., “Mitosis detection in breast cancer histology images with deep neural networks,” in [Medical Image Computing and Computer-Assisted Intervention MICCAI 2013], Lecture Notes in Computer Science 8150, 411–418, Springer Berlin Heidelberg (2013).
54. Cruz-Roa, Angel, et al. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks." Medical Imaging 2014: Digital Pathology. Vol. 9041. SPIE, 2014. 39 .
55. Janowczyk, Andrew, and Anant Madabhushi. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases." Journal of pathology informatics 7.1 (2016): 29.
56. A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, “Deep convolutional neural networks for breast cancer histology image analysis,” in International Conferen.
57. S. Kwok, “Multiclass classification of breast cancer in whole-slide images,” in International Conference Image Analysis and Recognition. Springer, 2018, pp. 931–940.
58. Y. S. Vang, Z. Chen, and X. Xie, “Deep learning framework for multi-class breast cancer histology image classification,” in International Conference Image Analysis and Recognition. Springer, 2018, pp. 914– 922.
59. A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab, “Structure-preserved color normalization for

histological images," in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE, 2015, pp. 1012–1015.

60. W. Nawaz, S. Ahmed, A. Tahir, and H. A. Khan, "Classification of breast cancer histology images using alexnet," in International Conference Image Analysis and Recognition. Springer, 2018, pp. 869–876
61. Kaggle, "Invasive ductal carcinoma dataset," Accessed on: 2019, https://www.kaggle.com/paultimothymooney/ breast-histopathology-images.
62. Chatterjee, Chandra Churh, and Gopal Krishna. "A novel method for IDC prediction in breast cancer histopathology images using deep residual neural networks." 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT). IEEE, 2019.
63. Boltaevich, Muminov Bakhodir. "Estimation affects of formats and resizing process to the accuracy of convolutional neural network." *2019 International Conference on Information Science and Communications Technologies (ICISCT)*. IEEE, 2019.
64. Kassani, S. H., & Kassani, P. H. (2019). "A comparative study of deep learning architectures on melanoma detection." Tissue and Cell, 58, pages 76–83.
65. Ronald A Rensink. The dynamic representation of scenes. Visual cognition, 7(1-3): 17–42, 2000.
66. Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus driven attention in the brain. Nature reviews neuroscience, 3(3):201–215, 2002.
67. Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. Neurocomputing, 452:48–62, 2021.
68. https://blog.paperspace.com/image-classification-with-attention/
69. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1409.0473.
70. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017
71. http://jalammar.github.io/illustrated-transformer/
72. https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0
73. Keles, Feyza Duman, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. "On the computational complexity of self-attention." *International Conference on Algorithmic Learning Theory*. PMLR, 2023.

74. M. Guo, J. Cai, Z. Liu, T. Mu, R. R. Martin, and S. Hu, “PCT: point cloud transformer,” Comput. Vis. Media, vol. 7, no. 2, pp. 187–199, 2021.
75. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker et al., “Attention gated networks: Learning to leverage salient regions in medical images,” Medical image analysis, 53, pp.197-207, 2019.
76. J. Son, W. Bae, S. Kim, S. J. Park, K. H. Jung, “Classification of Findings with Localized Lesions in Fundoscopic Images Using a Regionally Guided CNN,” In Computational Pathology and Ophthalmic Medical Image Analysis, pp. 176-184, 2018, Springer, Cham
77. https://www.edge-ai-vision.com/2022/04/multilayer-perceptrons-mlp-in-computer-vision/