Classifying Breast Cancer Images Using Vision Transformers

Zawadul Kafi Nahee
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
zawadul.kafi.nahee@g.bracu.ac.bd

Sabrina Rahman Mazumder
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
sabrina.rahman.mazumder@g.bracu.ac.bd

Tamzeedur Rahman Prithul
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
tamzeedur.rahman.prithul@g.bracu.ac.bd

Md Sabbir Hossain
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

Humaion Kabir Mehedi
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Md. Mustakin Alam
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Computer-aided diagnostic systems for the identification and classification of breast cancer have been developed using a variety of imaging modalities in conjunction with machine learning (ML) models. Performance effectiveness with ML-based machine diagnostic methods has advanced to something like a different extreme that also is comparable to actual specialists, thanks to fully convolutional networking (CNN) algorithms, which also have grown in latest years. CNN models are computational intelligence models. The investigation of screening for breast cancer has employed a variety of CNN models, notably supervised learning being more. Our Vision Transformer serves as the basis for something like the semi-supervised learning approach that we provide throughout this paper (ViT). As it has not been widely used in diagnosis of breast cancer, this ViT approach has been shown to outperform CNN models across numerous classification assessments. The recommended method offers a customized semi-supervised learning strategy that incorporates supervised as well as consistency training to boost the model's resilience. The approach also uses an adaptive token sampling strategy, which allows it to pick out the most crucial symbols from of the source image before testing, in particular to significantly enhance productivity. Researchers evaluate our methods on different datasets comprising contain pictures from histological as well as ultrasonography.**

Keywords: Semi-supervised learning, Detection of breast cancer, Vision Transformers, Samples of adaptive tokens, Data improvement, Diagnosis.

I. Introduction

The most prevalent malignancy among women is breast cancer. According to the World Cancer Research Fund's 2020 report, as there was over two million new cases of breast cancer in 2018[1]. These alarming figures emphasize the need of appropriately utilizing current technical breakthroughs to carry out effective breast cancer identification in its early stages. In order to create a software diagnosis method towards BC recognition that is greater efficient[3], a current advancement when it comes to artificial intelligence (AI), examines the use uses of deep learning techniques in a variety of medical fields is particularly encouraging. For the identification and diagnosis of breast cancer, a range of imaging methods can be utilized, including as X-rays (mammograms), ultrasound (sonography), thermography, magnetic

resonance imaging (MRI), and histopathological imaging[2]. In the process of diagnosing breast cancer, ultrasound has been an extensively used, inexpensive, imaging technique that is non-invasive and nuclear. Histological investigation typically follows it. Afterward employs biopsy procedures to obtain specimens of tissues and cells have been collected, put along a plate, stained, and examined under a microscope. The criterion for treatment is now histopathological, practically every cancer category with a high degree of certainty[6]. Despite the use of multiple imaging modalities, a visual inspection by radiologists or pathologists is still necessary, which takes time and needs a high level of radiological/pathological skill. Additionally, numerous studies have demonstrated that there is a significant amount when the same batch of pictures are taken, of cross – functional and cross inconsistency are interpreted by various specialists. In order to improve clinical decision-making, a more accurate diagnostic result might be produced by an AI-powered system, eliminating the evaluation disagreement brought on by varied experiences, analytical approaches, and expertise among humans[7]. In the health care sector, recent developments in AI, particularly throughout deep learning, have received a lot of attention. The use of deep learning for breast cancer identification had been used in a growing variety of use scenarios. First, the learning models, which include deep CNN frameworks already in existence, new CNNs, and hybrid versions that include a CNN, most commonly utilize convolutional neural networks (CNNs)[9]. In spite of the success methods for categorization depending on CNN, new developments have seen the emergence of the Vision Transformer (ViT), a unique vision model that has been proved to be even more precise in numerous open baselines. Few researchers have looked into the use of ViT throughout breast cancer, as well as its potential in this field has not been completely examined[4]. Second, the majority of current studies use supervised learning, which necessitates complete annotation of each image sample in the dataset. Semi-supervised learning incorporates a greater data set with no labels throughout training while only requiring based on a condensed sample of the training data[8]. Semi-supervised learning can significantly lessen the work required for annotating. Semi-supervised learning hasn't, however, been heavily utilized in recent breast cancer recognition research. Through the use of an adaptive token sampling method, the actual ViT technology is competent in a dynamic select the much more important picture tokens. In addition, we offer a unique consistency training approach to combine image enhancement using supervised and unsupervised knowledge acquisition. The dataset of the Breast Cancer Histopathological Image Recognition (BreakHis) set of data were used to validate the suggested technique. To exploit the unlabeled data, previous works employed semi-supervised learning.

II. Literature Review

These section describes previous study in various areas, covering semi-supervised learning used in biomedical image categorization as well as DNN-based breast cancer screening techniques.

The quantity of training set needed for such a totally supervised learning method can be decreased by the use of semi-supervised learning. This might be a lengthy process to collect a data item inside this biomedical sector, particularly in the area of disease research, where it might take months to assess a patient's ultimate condition[4]. Research that used a Semi-Supervised Support Vector Machine (S3VM) with specially created features for BC identification was published around 2016[10]. Around 2020, scientists employed PatchCamelyon-level artificial labeling to diagnose histopathology by identifying cancer cells that had spread to other organs[7]. Around 2022, several researchers carried developed genes utterance fate predicting for cancer recurrence using minimal separation, an semi-supervised learning technique[3]. Scientists created an SSL framework in 2010 which implements just several teaching by fusing an alignment system with such a human brain, greatly enhancing the model's learning capacity with less training data[5]. A few further uses for semi-supervised learning are colorectal cancer detection [11], skin cancer scare, and bladder cancer classification. To the state of the art, no studies before ours have investigated consistency training for breast cancer detection, thus our study intends to fix these problems.

Again, for categorization of breast tumors upon ultrasound as well as histopathological pictures, several pre-existing and bespoke deep CNN systems have been applied. Deep dnns, like CNNs, may learn racially discriminatory patterns with mechanically feature sets to depict an imagery sample[7], as opposed to tool ml algorithms that demand side features[16]. In order to merge CNN and CapsNet[4] to feature space fusion or improved routing, Chen et al[2]. created FE-BkCapsNet. To solve the convergence problem during training, Hamed et al. (2020) suggested combining the spatial properties of a CNN with the spectral information of a wavelet based[6]. New training procedures also been created in addition to the model enhancements.

III. Methodology

A. Dataset

The dataset of Histopathological imaging categorization for breast cancer (BreakHis)[10], that also represent non-invasive and invasive breast cancer detection approaches, According to the dataset, they are analyzed to validate the proposed method. However, by utilizing this dataset, we can train and test our model using photos from such a range of resources, that could be employed to assess a model's resilience.

1) BreakHis dataset: The BreakHis data consists 7,909 photos taken at absurdly low scrutiny.The dataset obtained breast tumor tissue from 82 women, consisting 2,480 benign as well as 5,429 malignant specimens. There still are four magnification levels in these shots. 40, 100, 200, and 400 to be specific. The samples are all 700 x 460 pixels, 3-channel RGB, and 8-bit deep in every channel, and all of them are kept in PNG format. Any malignancies criteria, such as meiosis,

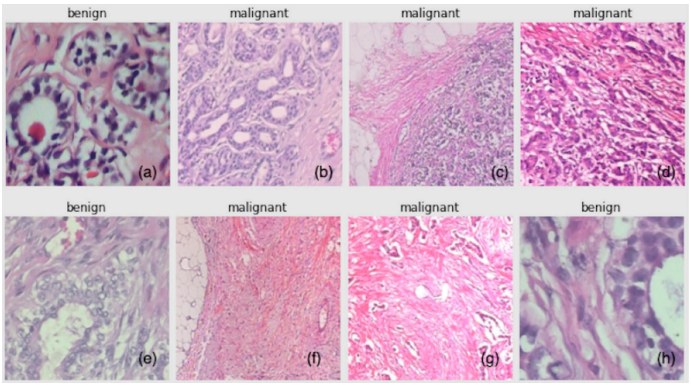
disruption of the basal cells, metastasis, etc., are just not in a sample that is fibrotic benign. In other phrases, adenomas swell and just don't expand. The malignant ones, on the other hand, have locally invasive lesions that can damage nearby structures and cause metastasis to other parts of the body. Table 1 shows our BreakHis sample statistics.

TABLE 1 shows our BreakHis sample statistics.

Magnification	Benign	Malignant	Total
x40	665	1,470	1,595
x100	674	1,477	2,051
x200	673	1,380	2,413
x400	585	1,732	1,880
Total	2,880	5,329	7,109
Patients	28	54	82

The breast tissue slides are digitally taken using a powerful image camera. Several recovered slides been colored with eosin and hematoxylin. These specimens are extracted by open surgical biopsy (SOB), as that is following by preprocessing for histopathological study and branding via physicians out from PD Research lab. The usual paraffin process, that is regularly utilized in healthcare practice, being used to prepare the tissues for this analysis. The basic objective to maintain the fundamental cellular structure as well as molecular structure so that it can be studied as seen using a fluorescence microscopy in undamaged state. After staining, the anatomopathologists use a microscope to visually inspect the tissue samples and assess whether each slide contains any malignant lesions. Each case's ultimate diagnosis is determined by skilled pathologists, and it is then supported by other examinations like immunohistochemistry analysis. Figure 2 displays a collection of samples from the BreakHis dataset, all of which are malignant except for the samples in subfigures (a), (e), and (h).

FIGURE 1 BreakHis samples: (A,E,H) are benign, and (B–D,F,G) are malignant.



B. The Learning Framework in Brief

The proposed method's overall workflow is depicted in Figure 3. The ViT must be trained as the primary model. The training process is divided into two steps: supervised training and consistency training. The former seeks to enhance the model's capacity for prediction, while the latter enhances its capacity for generalization. Through an end-to-end training process, both components are brought together. It should be noted that both training components share the ViT settings. Additionally, three different loss types are coupled to direct the gradient descent optimization of the neural network. The information below covers the training specifics.

C. Transformer

Throughout to address this sequentially modeled issues, this transformer is just a sort of neural method that employs a classification model to mined that store the meaningful linkages and semantics between it given tokens[11]. A inverter really does have the advantage of multithreading over cyclical architectures including prolonged poor memory (LSTM) the repetitive closed unit inside that words moving through its own construction may be consumed autonomously so instead of sequentially[12]. The inverter considerably exceeded expectations when used for language processing in NLP (NLP). My present investigation is one of the significant AI studies in the last five years since it also examines a lot of NLP works where its transformer was employed and produced a lot of pre-training approaches.

Each transformer seems to have an encoder-decoder design. This encoder component is created by stacking transformer encoders, while a decoded panel is created by stacking transformers. There is a consciousness function across every transformer encoding. Because to reflect the meaning interaction between of inputs tokens, there are several attentive units on a level. In that instance, awareness brain calculates how any token affects (pays attention to) one and other tokens using a tensor of scores. When results among these awareness units are aggregated, adjusted, and layered in a nutrient manner to create a library of Generative models are the product of the present encoder. The succeeding encoder continues the process using the embeddings produced by the preceding encoder as input. This feed-forward layer, decoders publicity layer, and non-linear and non self-attention layer seem to be the three layers to compensate a hybrid decoder, throughout contrasts. At each time step, a transistor decoder learns the embeddings itself from the former decoder, which would have represented the output of the decoder function at the previous stage for the first decoder. As to achieve the accurate results, this coder then feeds the user input through a stack of decoders, a linearity layer, as well as a max pooling layer.

D. Vision Transformer

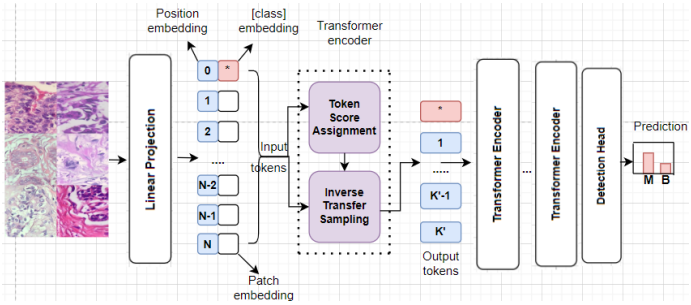
After discovering how efficiently a transformer performs in NLP tasks, academics began to consider its possibilities in object recognition. The ViT was one of the first initiatives. The ViT maintains the construction of the previous transducer, but the input has undergone the following changes. An picture is divided up into a number of distinct image patches in order to meet the input requirements of such transformers. A fully linked layers that experiences a linear alteration is fundamentally what the so-called image patch learning

algorithm is all about. Additionally, given a sizeable input image of H × W × C is split into N patches (i.e., tokens), each of size P × P × C, then we can determine that N = $\frac{HW}{P^2}$. Subsequently, a rectangle of dimensions D is created by distributing each patches. Each source was therefore converted into something like a 2D net of length N * D. All classify metadata is also encoded by adding a special symbol to the starting stance of the set of time series. Such technique is commonly used in different pre-training techniques, for instance the Bidirectional Encoder Characterizations of Transformers (BERT)[13], it adds a positional encode matrix toward each patching immersing to construct a token encapsulating in the initial layer of own transformer encodes.

E. Adaptive Token Sampler

Because, since the cost of calculating increases quadratically with the number of tokens, any ViT is difficult to solve. CNNs use several pooling methods to lower the resolution inside the chain. Unfortunately, employing pooling there in ViT is just not practical since the tokens are permutation invariant. To cut down on processing costs, we hence use an adaptive token sampler, a method that enables the algorithm may proactively select important symbols out from supplied strings. This network infrastructure of ViT is shown in Figure 2.

FIGURE 2 ViT-architectural, ViT's design. Token value assigning and inverted shift polling may both be carried out by the ViT program, which can be included through every transformer square. So class label may be increased, or the data size efficiently decreased by the ViT by determining the most important bits for feed along to the succeeding tiers.



F. Semi-supervised Learning

In order to increase a model's resilience, this semi-supervised learning training strategy investigates simultaneously labeled as well as unlabeled data. Additionally, semi-supervised learning is a well-liked technique in situations when there are few training examples available due to expensive labeling prices. Our smoother principle, is used in this work along with many semi-supervised learning instructional strategies, states that comparable pictures should really be classified as belonging to almost the same classification[5].

IV. Results

A deep learning technology utilized in this investigation, PyTorch 1.8.0, has been coded in Python 3.6.10. On a workstation with Windows 10 installed, an i7-11900k processor, and then an Nvidia GTX2080 Super 12G graphics card, most test was conducted.

A. Evaluation Metrics

Because the courses across both samples are unbalanced, accuracy (Acc) alone is insufficient to accurately represent an effectiveness of the algorithm. Consequently, in complement to ACC, we also employ F1 scores, precision (Pre), as well as recall (Rec) total score to evaluate performance. Throughout Eqs 10–13, following markers are provided

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$...(10)

Precision = $\frac{TP}{TP+FP}$...(11)

Recall = $\frac{TP}{TP+FN}$...(12)

F1 score = $2 * \frac{Precision*Recall}{Precision+Recall}$...(13)

in which the terms true positives, true negatives, false positives, as well as false negatives, consecutively, stand for the percentage of each. The frequency of false alerts is shown in Precision. That program may have less false alerts, the bigger sensitivity precision. Recall displays the number of overlooked positive cases during that time. In these other circumstances, smaller positive samples have already been overlooked in the stronger index Recall. F1 is the transfer function of Precision as well as Recall, making it a better statistic for such a categorization job with either an unbalanced dataset over Accuracy.

B. Preconditions

For starting points with this investigation, four models—the VGG16, ResNet101, DenseNet201, as well as ViT are being selected. These four approaches have been successfully applied in several picture classification methods and therefore have provided reliable starting points.

This same VGG16 networks consists made up of five chunks that are separated by something like a deep network, allowing automatically resize filtering. So every unit has two or three fully connected layers, enabling extraction of features. Three hidden layers, one of which is a soft-max layer, are added after the final frame to provide standardized vectors that serve also as the outcome of something like the guess. Another foundation for creating a deep but precise network seems to be the wide usage of tiny (3 * 3) convolution layers inside this VGG machine learning model.

ResNet's version, DenseNet, differs from it in two ways. First, the level product and indeed the bypass inputs within every block are combined by DenseNet using a concatenation rather than a total (used by ResNet). Across many blocks are separated by a transitions layer introduced by DenseNet. To manage the number of hidden layers, each crossover layer consists of a (1 * 1) convolutional layer as well as with average pooling level throughout a 2 duration[15].

Our ResNet neural structure builds a series of pooling layers, which together makes it easier to teach an internal

representation by channeling its intake into its output through a relatively brief link. Which makes it simple to establish an identifying function, which makes it possible could train a structure containing additional layers multiple efficiently before experiencing decreasing returns[16]. ResNet101 has a combination number of 101 layers but is composed of repeating activation functions, a dense layer, as well as a softmax layer.

Section II covers the discussed ViT.

C. Training Framework

Table 3 displays the primary hyper-parameters utilized during training. Adam was chosen as the optimizer, and his learning rate was 2e-5. To avoid the denominator from becoming 0, we set eps = 1e-08. The batch size was set at 64. The binary cross entropy with logits employed as the loss function.

TABLE 3 Training Framework.

Hyperparameter	Value
Gain	2e-5
Eps	1e-12
Array Volume	32
Epochs	300
Intake Image Size	224 * 224
ViT Tokens	[256, 128, 64, 32, 16, 8]

300 epochs were required to train each model that was assessed. Each input picture was divided into 16 patches and re-scaled to a fixed size of 256 x 256 for the ViT. Six encoders make up the ViT model employed in the study. The default option from the ViT's original report was used in the ViT method, which stored 256, 128, 64, 32, 16, and 8 tokens at every layer.Several variables got determined using study evidence. It needs to be noted that perhaps the investigators explored with different token sample groups in addition here to default setting, but didn't see that the results significantly changed. This seems to be in order to make it possible to adapt the very same model's sequential extracting capabilities for different input codecs.

Our training and test sets from both datasets are divided into them in the following proportions: 7:3. Further, dividing the training set into an 8:2 ratio, 20% of the data are considered as unlabeled data utilized for ViT, while the remaining 80% are used during the supervised training to create an ViT architecture.

D. Results

Overall comparative evaluation of the suggested technique and indeed the selected baselines is shown in Table 4. To assess the effectiveness of the adaptive token sampler and consistency training, an ablation research has also been carried out. To create the ViT model, we took the ViT as a basic model as well as integrated the adaptive token sampler as well as consistency training. Four metrics—Acc, Pre, Rec, and F1—defined in Section III have been presented for each examined model. We offer the following explanation of the results.

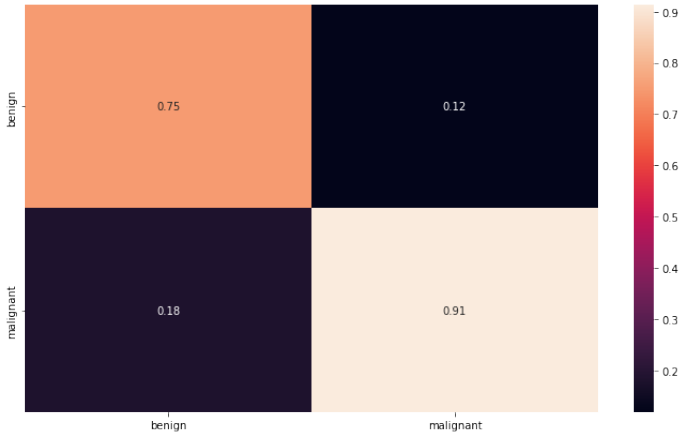
TABLE 4 Results on BreakHis.

Model	Acc	Pre	Rec	F1
VGG19	95.75	95.39	94.68	95.21
ResNet101	94.64	94.32	96.49	94.18
DenseNet201	96.74	92.82	95.10	96.56
ViT	**97.79**	**97.32**	**97.44**	**97.63**

The outcomes of the verified algorithms on BreakHis are displayed in Table 5. These remarks are addressed in this section.

DenseNet201 displays the greatest Acc of the three benchmark algorithms (96.74%), whereas VGG19 exhibits the maximum F1 of 95.21%. With such an Acc at 97.79%, a Pre at 97.32%, overall Rec at 97.44%, as well as an F1 of 97.63%, ViT, but in the other end, performs the best across all four measures. This outcome demonstrates that ViT may consistently enhance performance across both databases and seems to be a potential tactic to increase a model's generalizability[14].

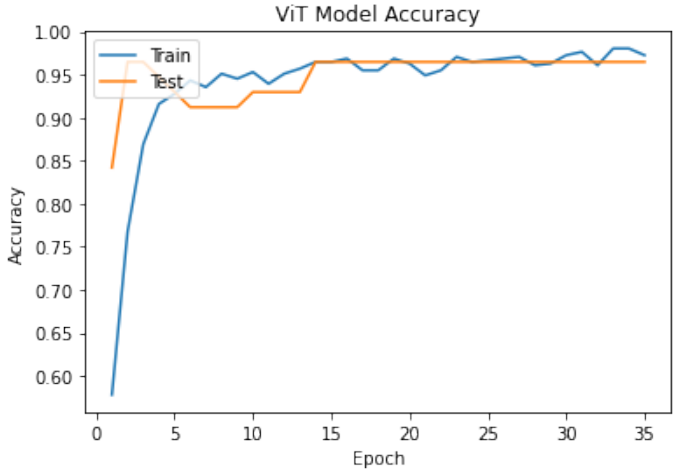
FIGURE 3 ViT model confusion matrix indicates effectiveness.



On evaluating a classifier's efficacy, a further N x N matrix dubbed a confusion matrix has been used, with N is a number of specified classes. In some kind of matrix, the actual objective values are compared to what the machine training system had predicted. This provides us with a profound comprehension of the efficacy of our classification system, as well as the kinds of errors it is making. Confusion matrices reveal discrepancies between the predicted and actual numbers. Their response TN, which stands for True Negative, discloses the number of accurately recognized negatively classified instances. In a manner similar to this, TP stands for True Positive and refers to the total number of correctly detected positive instances. This same proportion of true negative occurrences that had been wrongly particularly pronounced is known as both the false positive benefit (FP), meanwhile the percentage of true positive events that had been incorrectly classified as negatives are known as

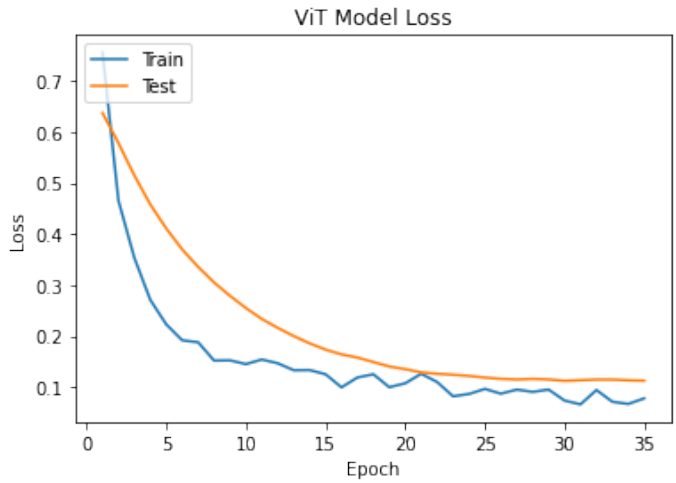
both the false negative value (FN). Nowadays, among the categorization metrics that is used most often is precision. ViT simulation deficit in Figure 4 suggests efficacy. Learning biases are a commonly utilized medical diagnostic in deep learning with regard to approaches which acquire information gradually using a training of examples. This method may be evaluated just on a training sample and followed by a freeze verify dataset regularly after modifications during training, meanwhile diagrams of the quality metrics may be constructed to illustrate curves.

FIGURE 4 ViT model accuracy graph indicates effectiveness.



Regarding techniques who gain knowledge progressively from such a training dataset, learning gradients are an often used medical diagnosis in deep learning. Every after updates throughout training, this algorithm may be tested upon that training dataset and then a hold-out verification dataset, while graphs of the performance measures can be made to display curves.

FIGURE 5 ViT model loss graph indicates effectiveness.



V. Conclusion

This work introduces ViT, a ViT model that was enhanced using adaptive token sampler and trained using consistency training. The suggested model has been evaluated on the breast cancer neuroimaging dataset and has outperformed three standard CNN algorithms in comparison. The outcomes have shown how effective ViT is. Both the first and the second combine supervised and unsupervised training to increase the model's capacity to generalize. The former enables the autoencoder to pinpoint key regions of focus that offer interesting trends for something like the classification job. The validated data from the suggested model make it a reliable standard for further studies.

The results of this study include a number of noteworthy ones. Our findings show that, when compared to its CNN rivals, the previous ViT framework does not exhibit improved performance. The ViT is somewhat poorer, but still similar to the BreakHis datasets. This could be as a result of the breast cancer identification job, where the pictures might include delicate patterns that are challenging to notice even using the ViT's conscience system.

A major functional component of a computer-aided diagnostic model for breast cancer diagnosis might be the hypothesized ViT approach. It has two advantages. In the beginning, the adaptive token sampler feature enables the system to display the most instructive picture patches, which can assist doctors in swiftly identifying the crucial regions for accurate and customized diagnosis. Second, once new photos are ready, the computer-aided diagnostic system's backside may be readily changed to function as a learnable skill engine. Due to the consistency training's semi-supervised nature, only a small fraction of such recently contributed data has to be tagged, thus lowering the price of labeling.

Several subsequent approaches can indeed be taken to expand the suggested technique. Any aggregation of the 2 or functionality aggregated might be a different model construction approach that may combine the advantages of both neural architectures. Initially, we contrasted CNN models with the ViT primarily. Provided that such CNN and the ViT have substantially different foundational models, the CNN requires numerous filters to capture multiscale features while the ViT examines the semantic relationships among each duo of items. Combining the two might also result in better performance than using either model alone. A number of co breast cancer neuroimaging dataset with multidimensional extracted features might be used to test the suggested technique. In conclusion, the use of vision transformer (ViT) for classifying breast cancer images is a promising development in the field of medical imaging. The advanced machine learning technique has demonstrated exceptional accuracy and efficiency in detecting and diagnosing breast cancer. The ability of vision transformer (ViT) to process and analyze vast amounts of data makes it a valuable tool for doctors and researchers working to combat this disease. As research in this area continues to advance, it is likely that the use of vision transformer will become

increasingly prevalent in the field of breast cancer diagnosis and treatment. The effectiveness of vision transformer in classifying breast cancer images highlights the potential of this technology to revolutionize the way we approach medical imaging and breast cancer diagnosis.

REFERENCES

[1] Gheflati, B. and Rivaz, H., 2022, July. Vision transformers for classification of breast ultrasound images. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (pp. 480-483). IEEE.

[2] Chen, X., Zhang, K., Abdoli, N., Gilley, P.W., Wang, X., Liu, H., Zheng, B. and Qiu, Y., 2022. Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms. Diagnostics, 12(7), p.1549.

[3] Tummala, S., Kim, J. and Kadry, S., 2022. BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers. Mathematics, 10(21), p.4109.

[4] Gheflati, B., and Rivaz, H. (2021). Vision transformer for classification of breast ultrasound images. arXiv Prepr. arXiv:2110.14731.

[5] Chen, K., and Wang, S. (2010). Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions. IEEE Trans. Pattern Anal. Mach. Intell. 33, 129–143. doi:10.1109/TPAMI.2010.92

[6] Hamed, G., Marey, M. A. E. R., Amin, S. E. S., and Tolba, M. F. (2020). "Deep learning in breast cancer detection and classification," in The International Conference on Artificial Intelligence and Computer Vision (Berlin, Germany:Springer), 322–333.

[7] Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. Adv. Neural Inf. Process. Syst. 33, 6256–6268.

[8] Van Engelen, J. E., and Hoos, H. H. (2020). A survey on semi-supervised learning. Mach. Learn. 109, 373–440. doi:10.1007/s10994-019-05855-6

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

[10] Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016). A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. 63, 1455–1462. doi:10.1109/TBME.2015.2496264

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Adv. neural Inf. Process. Syst. 30.

[12] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv Prepr. arXiv: 1412.3555

[13] Fayyaz, M., Kouhpayegani, S. A., Jafari, F. R., Sommerlade, E., Joze, H. R. V., Pirsiavash, H., et al. (2021). Ats: Adaptive token sampling for efficient vision transformers. arXiv:2111.15667 [cs]

[14] Moon, W. K., Lee, Y. W., Ke, H. H., Lee, S. H., Huang, C. S., Chang, R. F., et al. (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput. Methods Programs Biomed. 190, 105361. doi:10.1016/j.cmpb.2020.105361

[15] Singh, D., and Singh, A. K. (2020). Role of image thermography in early breast cancer detection-Past, present and future. Comput. Methods Programs Biomed. 183, 105074. doi:10.1016/j.cmpb.2019.105074

[16] Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., and Abdel-Mottaleb, M. (2021). Convolutional neural networks for breast cancer detection in mammography: A survey. Comput. Biol. Med. 131, 104248. doi:10.1016/j. compbiomed.2021.104248