International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 11, Issue 2 (2025)
https://doi.org/10.31436/ijpcc.v11i2.558

Detection of Errors in Bitewing X-Ray Images Using Deep Learning

[1]Aiman Syahmi Bin Ahmad Sabri, [1]Akeem Olowolayemo, [2]Ahmad Badruddin Ghazali, [3]Ibrahim Muhammad, [4]Fatimoh Damola Saliu-Olaojo

[1]Department of Computer Science, KICT, International Islamic University Malaysia, Kuala Lumpur, Malaysia.
[2]Department of Oral Maxillofacial Surgery and Oral Diagnosis, Faculty of Dentistry, International Islamic University Malaysia, Kuantan, Malaysia
[3]Computer Engineering Department, School of Science and Engineering, Alhikma Polytechnic Karu, Nasarawa State, Nigeria.
[4]Department of Computer Science, Faculty of Natural Science, First Technical University, Ibadan, Nigeria.

*Corresponding author: akeem@iium.edu.my
(Received: 15th January 2025; Accepted: 18th May, 2025; Published on-line: 30th July, 2025)

Abstract— Quality assurance (QA) is a process put in place in the hospital to guarantee ideal diagnostic image quality with minimum danger to patients. It entails frequent quality control checks, preventive support procedures, authoritative approaches, and planning. The process of acquiring quality images, especially for radiography students and trainees, requires a steep learning curve. This study proposes deep learning models that may serve as a guide to ensure proper images are captured and help improve the quality assurance process. The models are intended to determine that the images captured are optimal by ensuring adequate precautions in the capturing process, thereby automatically identifying and correcting any mistakes or issues in the quality or interpretation of the image. This study acquired 4955 radiographs that have been labeled by dental experts. Four deep learning models, specifically CNN, AlexNet, RestNet-50, and ViTs have developed with respective accuracies of 78.98%, 24.84%, 78.03%, and 81.34%. The performance results show that deep learning models have the potential to be utilized to assist dental practitioners in error detection and quality assurance.

Keywords— Bitewing radiography, Bitewing radiography Error, machine learning, convolutional neural network (CNN), AlexNet, Residual Neural Network-50 (ResNet-50), Tensor, Visual Transformers (ViTs), Image Classification.

I. Introduction

Artificial intelligence (AI) has witnessed an exciting surge in advancement in recent times. Machines and equipment are evolving rapidly to make human tasks and chores easier. Nonetheless, a prevalent design issue remains in their ability to consistently produce high-quality output. The reason behind this is that achieving a quality output fundamentally relies on domain-specific solutions, requiring concerted efforts. Complex machines entail numerous quality assurance procedures and demand technical expertise. This indicates that human effort and time continue to hold a considerable level of importance. This is particularly relevant in the application of machine learning in the medical field, such as radiography. Radiography is a medical technique that encompasses the creation of diagnostic images like X-ray, ultrasound, Computerized Tomography(CT) scan, and Magnetic Resonance Imaging (MRI). Analyzing radiographic images for diagnosis where individuals had to endure over a month-long wait for their X-ray results to be processed may not be desirable for an efficient health care system. This delay can be attributed to the time-consuming nature of imaging examinations and procedures that heavily rely on human involvement. In radiographic diagnosis, there exists considerable potential for errors, due to human factors as well as defective machines.

Anomalies of the human body are captured by imaging techniques. In order to diagnose, and plan therapy for the anomalies, it is necessary to comprehend the collected images. Medical experts with expertise often interpret medical images. Nevertheless, the efficiency of image interpretation carried out by qualified medical specialists is limited by the scarcity of human experts, their weariness, and the imprecise estimation processes associated with them. Errors in assessing radiographic images may lead to either a false positive or a false negative. In the case of a false positive, a patient may be subjected to emotional trauma, asked to undergo a life-changing procedure, incur an expensive financial loss, and a considerable waste of time. False negative, on the other hand, could lead to delayed treatment consequently increasing the chances of further complications which in turn would lead to more critical procedures and impact the patient's life. In this study, an attempt has been made to auto-detect and classify the common errors in bitewing X-ray images. The experiment was conducted to determine or detect that the images

captured through the process of radiography satisfy the required quality. This is to identify and correct any mistakes or issues in terms of the quality of the image and to ensure the quality assurance process of the image does not stray when conducting the radiography process. Deep learning classification algorithms based on the Convolutional Neurol Networks (CNNs), as well as other variants such as the Inception Neural Networks (InceptionNets), Residual Neural Networks (ResNets), and Vision Transformer (ViTs), have been employed in conducting the image classification of the quality of the X-ray images. The experiment aims to assist radiographers as well as medical students in increasing the quality of X-ray images. Additionally, previous data would be useful as a reference for medical students to avoid preventable errors when an X-ray image is being taken.

II. Related work

Researchers' interest in the application of deep learning to medical domains has been consistently growing in recent years. Deep neural networks, or DNNs, form the core of emerging artificial intelligence (AI) systems. The most established algorithm among various deep learning models especially for medical images is convolutional neural networks (CNNs), a class of artificial neural networks that has been a dominant method in computer vision tasks. Convolutional neural networks (CNNs) process shift-invariant input, such as images, by introducing convolutional layers and pooling layers ultimately linked to the fully connected layer for final classification.Applications of CNNs among radiology researchers have been published in areas such as lesion detection, disease classification, infected area segmentation, image reconstruction, and natural language processing [1]. Researchers using CNNs for medical imaging and radiology duties could potentially impact clinical radiologists' work. This covers CNN's opportunities and potential future paths while concentrating on the fundamental ideas of AI and how they apply to different radiology applications [2].

Convolutional neural networks (CNNs) are effective tools for image understanding. They have been shown to outperform human experts in many image analysis and understanding tasks [3]. The ultimate goal is to encourage academics studying medical image interpretation to utilize CNNs extensively for diagnosis and research. Many researchers have designed automated systems for extracting fundamental features from images [4]. Convolutional Neural Network (CNN) is a popular deep learning method for computer vision applications [7]. The human ability to recognize objects visually served as the inspiration for this deep learning system [8]. One of the algorithms created to help academics and researchers with classification issues is CNN where it makes the possible to use images as input, which moves artificial intelligence technology one step closer to mimicking humans' use of sight, a different sense, to understand their environment. Only words and numbers could be fed into its older algorithms. CNN uses artificial neurons instead of actual ones to detect objects in a similar manner to how people use their own brain neurons [9]. To enable the computer to distinguish between each pixel in an image and produce the right result, CNN would process and extract from an image through multiple levels of processing.

Another recent model is the Transformer, which mainly utilizes the self-attention mechanism, to extract intrinsic features. Often, this is mainly utilized in large language models showing great potential for extensive use in AI applications [5]. When the Transformer model was initially used, it significantly improved natural language processing (NLP) tasks. For machine translation and English constituency parsing tasks, for instance, initially proposed the transformer model based on the attention mechanism. A novel language representation approach, Bidirectional Encoder Representations from Transformers (BERT) was proposed to pre-trains a transformer on unlabelled text by considering the bidirectional nature of each word's context [6]. Images are thought to be more challenging for generative modeling than text because they incorporate additional dimensions, noise, and duplicate modality. The transformer can be used as the backbone network for image categorization in addition to CNNs. Wu et al, substituted vision transformers for the last convolutional layer and used Residual Network (ResNet) as a practical baseline [Wu et al]. Convolutional layers are specifically used to extract low-level characteristics, which are then sent into the vision transformer. To arrange pixels into a limited number of visual tokens for the vision transformer, each of which represents a semantic concept in the picture, employing a tokenizer. The direct application of these visual tokens has been seen in picture categorization.

III. Methodology

The deep learning architectures that will be employed in this study are CNN and variants such as InceptionNets, DesNets and AlexNets.

A. Convolutional Layer

In this research, the dataset was acquired from IIUM Kuantan Medical Campus after approval from the university ethics committee. Subsequently, the dataset was manually extracted from the machines, while Dental experts assigned classes to the extracted images. This dataset is then prepared for the deep learning models. The features from the provided dataset were extracted using a conventional CNN architecture. The dental X-ray image dataset needs to be preprocessed before being fed into the models, as will be

https://doi.org/10.31436/ijpcc.v11i2.558

covered in more detail subsequently. As can be seen in Figure 1, the CNN model consists of convolution layers, max-pooling, average pooling, and fully connected layers. The number of channels and fully connected layers (FC) as well as the filter size of the convolutional layers are indicated by the notation in each block. The ReLU function, max-pooling layer, average pooling, and flattened layer are indicated by additional block labels [10].

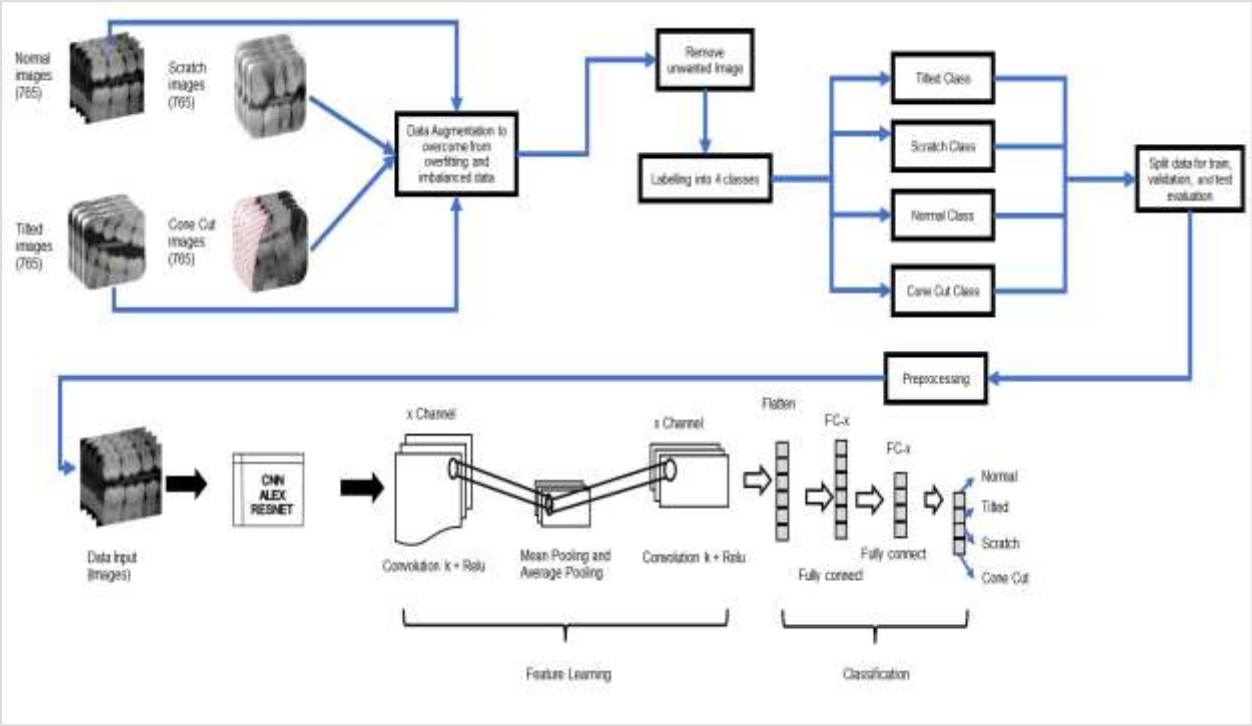

Fig. 1 Methodology of the project for CNN

Since images are inherently non-linear, the Rectified Linear Activation function (ReLU) is applied after each convolutional layer to introduce non-linearity into the model since it returns the same results regardless of whether they are positive or negative, this method also aids in speed (see Figure 2).

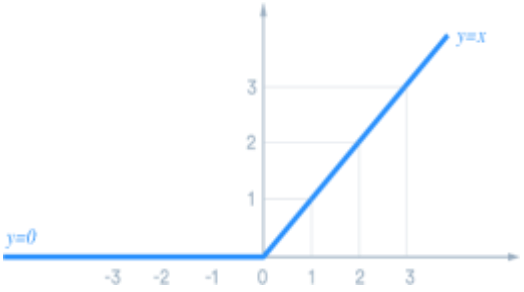

Fig. 2 ReLU Activation Function

B. *Pooling Layer*

Following input through the first convolutional layer, the feature map is shrunk by the pooling layer, which consists of a pool and a stride[11][12]. A pool will move across the feature map in accordance with the pooling approach in order to extract features. The pool action's both horizontal and verticals are determined by a stride.One of the several pooling layers in this model is max-pooling. Prior to traversing the flattened layer, each convolution's output will undergo max-pooling. A number of methods can be used to achieve pooling, for instance max, average, and min pooling, which produce the maximum, average and minimum values from the dental X-ray section that corresponds to the kernel, respectively. Similar to the convolutional layer before it, the pooling layer considers factors like stride and layer size..

C. *Fully Connected Layer*

The dental X-ray feature map is extracted using the global average pooling operation after completing the last module and before it reaches the fully connected layer. The network's final layer is the fully connected layer, which offers better classification performance than features retrieved from earlier layers. It is an illustration of a traditional neural network architecture, where a dense network is created by connecting every neuron in one layer to every other layer's neuron [13]. The fully connected layer

https://doi.org/10.31436/ijpcc.v11i2.558

receives the final convolution's output as input, flattens it, and then passes it through the fully connected network for classification. Before the inputs are fed into the fully connected layer, the flattening operations is necessary because the final convolution produces a dimensional matrix as its output. All of the matrix values will be converted into vectors in order to achieve flattening necessary to perform the SoftMax procedure before classification.

$$y = f(Wx + B) \quad (1)$$

In artificial neural networks, this is a standard equation. Wx is the dot product between the weight matrix W and the input vectorx[14]. The bias term b is added to the result of this dot product. Finally, the activation function f is applied element-wise to the result.

For this research, the input will be categorized into three previously described classes of dental X-rays based on the probability of the item in the classes [15][16].

The Transformer architecture was first created for natural language processing, but Vision Transformers, or ViTs, is a deep learning model that adapts it for computer vision tasks. It has drawn interest because it can perform competitively without the use of convolutional neural networks (CNNs) in image classification and other vision tasks.

Vision Transformer employs a technique known as self-attention [5]. Because it allows computers to comprehend the relationships between the various components of an image, it is also known as self-attention. Additionally, a self-aware computer can concentrate on distinct image patches and comprehend how they connect to create a complete picture. In essence, what Vision Transformers do is divide the image into smaller units known as patches, and by turning them into a series of tokens, the Transformer model is able to analyze and comprehend each component of the image independently.

D. Transformer Model

The Transformer network is intended to resemble the attention process. Instead of using visual cues like movie frames, words in a sentence, notes, or even individual pixels in an image, it uses attention to comprehend the order of information [5]. Therefore, even when components are far apart, Transformer networks are still able to capture the dependencies and relationships between them. For tasks like language understanding, transformer networks are particularly effective because of their capacity to capture long-range dependencies (see Figure 3).

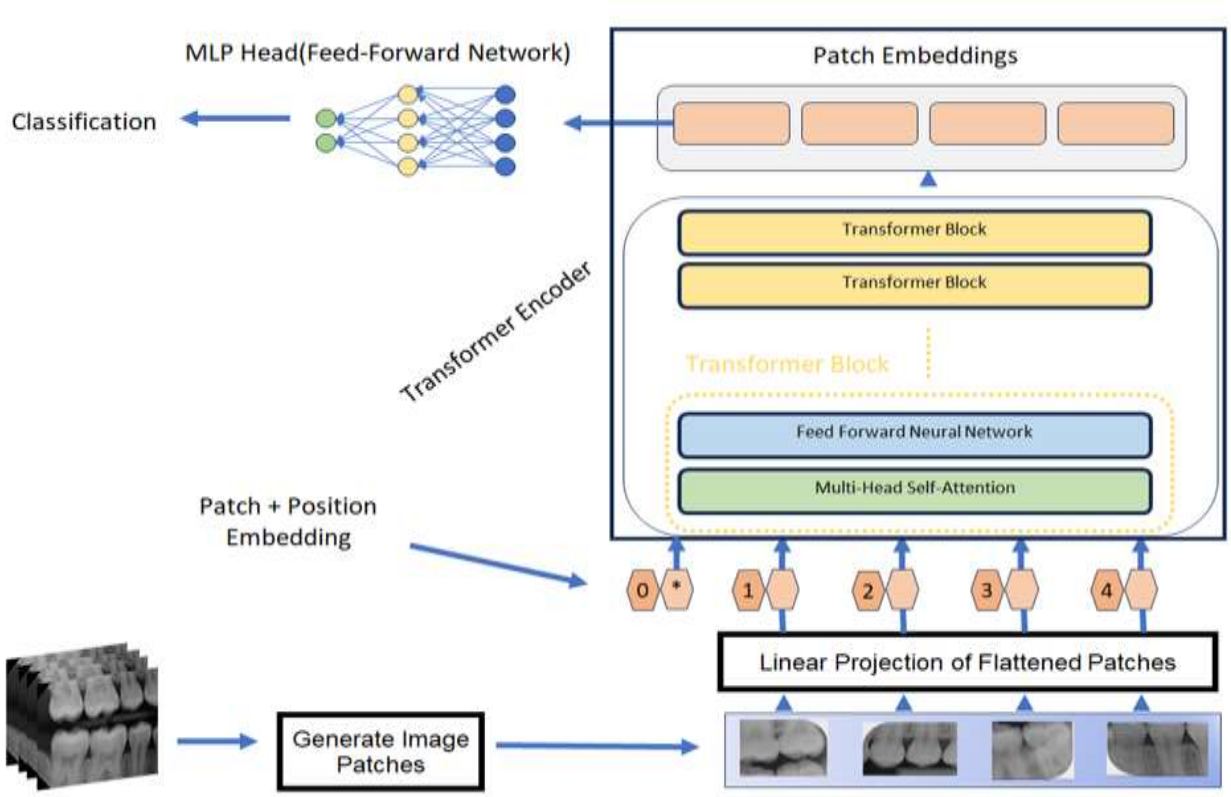

Fig. 3 Methodology of the project for Vision Transformer

https://doi.org/10.31436/ijpcc.v11i2.558

E. Linear Projection

Visual projection that is linear each 1-D vector will be converted into a lower dimensional vector by the transformer as it works on these flattened patches, while keeping the connections and significant elements. The two primary steps in linear projection are bias addition and weight matrix multiplication.

After completing these two stages, a vector is transformed into one with a lower dimensionality, or one with fewer elements or components than the original vector to extract key characteristics and record the most crucial data while eliminating the less crucial details. By removing noise and pointless variations from the data, dimensionality reduction actions can strengthen and narrow the focus of the vector representation to the most important features.

F. Position Embedding

Every image patch has position embedding added to it, which shows every location in the image. Data were fed into the transformer using positional encoding and with the help of this positional embedding, position information for every patch that is available were fed to the transformer vector, which is then fed to the subsequent layers in the vision transformer for additional processing.

G. Self-Attention Layer

The transformer encoder's first layer is the self-attention layer. Self-attention enables every patch to pay attention and learn from other patches. It allows the model to take the global context into account by capturing dependencies between the patches.

Understanding the relationship between the patches or tokens in an image is aided by the self-attention layer. It assists the model in determining which patches are related and how important to one another .

All of the patches in the image are taken by the self-attention, and each patch is assigned by three unique jobs that seek queries, functioning similarly to a patch searching for other patches to focus on, while the key functions similarly to a patch being examined by other patches, and valuefunctions similarly to the patch's details or information.

It computes a similarity score between every patch inside of it. The more closely related the patches are, the higher the score. These similarity scores are then applied to each patch to determine the relative importance of each word in the image. More attention will be paid to the patch with the higher score. Subsequently, the self-attention will aggregate the relevant patch values according to their respective attention weights andbuilds a new representation for every patch by assembling the relevant image's details.

H. Feed Forward Neural Network

The next in line after the self-attention layer is the feed forward neural network. Every patch's output is sent through the feed forward, which uses it to help identify intricate non-linear relationships between the patches.

I. Add & Norm (Residual Connection)

The add and norm layer, also referred to as the residual connection or skip connection, performs an element-by-element edition between the feed-forward or attention output and the output of the preceding layer. The original data from the previous layer, which aligns the model to learn and update the new information captured by the supplying layer, is preserved with this addition. By including this sublayer's output, the original input add layer gave the model a shortcut path for information flow and aided in the gradients' effective propagation during training. Thus, these layers are essential for managing information flow and maintaining training process stability.

J. Dataset

The bitewing radiographs data for the study was collected at thefaculty of dentistry, International Islamic University Malaysia, Kuantan by dental experts after receiving ethical approval from the committe. A total of 3060 bitewing radiographs were extracted and labelled into classes. The dataset contains 765 of each of the classes, namely cone cut, normal, scratch and tilted respectively. Samples of the different classes are shown in Figure 4.

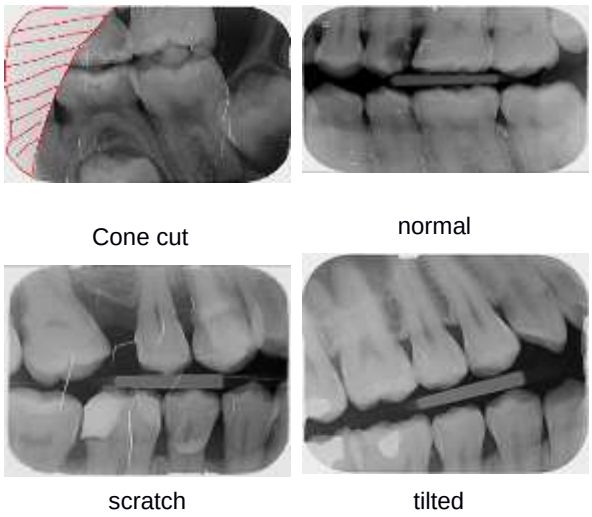

Fig. 4 Labelled radiograph

K. Preprocessing

The first stage in the CNN techniques to preprocess the image for classification [17].

https://doi.org/10.31436/ijpcc.v11i2.558

The training and test sets are split in a 80:20 ratio which consists of a total of 96 radiographs batches according to the size of batches.

The radiographs' channels are rescaled between 0 and 1 before the model is trained to standardize the input. To increase the pace of model training, they are reshaped to 224 by 224 pixels [18].

L. Training

Each of the models for CNNs architecture is being set for the same structure to compare the effectiveness with 30 epochs to avoid model overfitting. As for the batch size, it is set to 64 to maximize the speed of the device.

On the other hand, for the Vision Transformer model, it has a different architecture compared to CNN model. In ViTs, there is the need to measure input features set to 768 (16*16*3) and output features to classesnce the ViTs requires a high hardware specification, the utilization of transfer learning is necessary to avoid exceptions from occurring during training. The epochs are being set to 30 epochs to measure training efficiency. The batch size is setup to 64 to train the batches available. In the data loader, the number of workers has been fixed to a maximum of 4 to carry the full potential of the CPU and GPU available. The model of CNNs is being trained using the Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz processor (CPU) for the laptop specification. As for the ViTs is being trained using the Nvidia GeForce 16 series (GPU) with 4 gigabytes (GB) of Video Random Access Memory (VRAM) on Desktop. The CNNs model is built on Python v3.10.3 with Keras v2.10.0 and TensorFlow 2.10.0 as site backend. While ViTs, build on same Python version with Pytorch v2.1.2 along with CUDA v11.8 to utilize GPU.

M. CNNs architecture

CNN is the most common deep learning algorithm in image classification. Generally, it is a deep neural network model that consists of two parts, namely; image feature extraction and classification. The proposed baseline CNN has 3 layers of convolution with Rectified Linear Unit (ReLU) activation and pooling, which are inserted alternately. Multiple dropout layers of 0.2 and 0.5 dropout rates are implemented in between. Eventually, there are roughly 65 hundred thousand trainable parameters.

The design that has been recommended consists of two dense layer blocks, three pooling blocks, and three convolutional blocks, in that sequence. A few batch normalizations are being done to make matrix computation faster [19], [20].

On the other hand, Vision Transformer is a variation of natural language processing model that focuses on the self-attention layer where all the tokenization is arranged before patch embeddings. The utilization architecture consists of convolution projection, encoder blocks and linear heads.

The three CNN architectures are designed in such a way that the feature extraction part gradually pools the radiograph until it is an input of single-digit by single-digit pixels [21].

N. Hyperparameter settings

All the CNNs will be trained using the same hyperparameter settings. A few of them, such as the number of epochs and batch sizes, have been mentioned. Weights are initialized based on the default settings. The loss function used is Categorical Crossentropy. It measures the dissimilarity between the predicted probability distribution and the true distribution.

The loss function is optimized by Adam optimizer and set to have a learning rate of 0.001.

IV. Results

The results below will show the accomplishment that has been acquired after fitting the models between these CNN and ViTs models according to their limitations.

A. Base CNN

Based on Figure 5, we can conclude that the loss rate is minimum at most to 0.3 value rate while the starting loss rate count from 3.5 value rate when running under a total of 100 epochs. Meanwhile, the loss rate for the validation reduces by only a portion of the loss which is a 1.0 loss rate starting from a 1.5 value on 100 epochs.

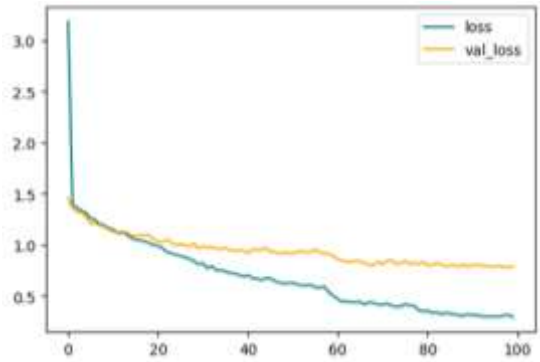

Fig. 5 Figure shows the loss rate of Base CNN

Based on Figure 6, the starting accuracy value is around 0.25 while decently ascending as the number of epochs runs. The accuracy is around 0.92 on 100 epochs for training, followed up by the validation accuracy rate of 0.79 plus.

Based on the prediction result in Figure 7, the number of actual labels that meet prediction for cone cut, normal, scratch, and tilted are 114, 133, 146, and 104 respectively. The most correctly predicted class is the scratch class.

https://doi.org/10.31436/ijpcc.v11i2.558

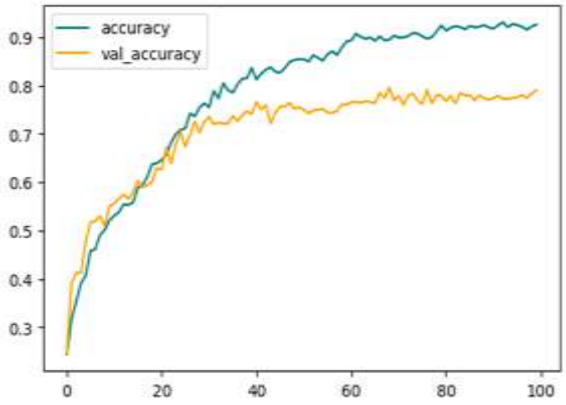

Fig. 6 The accuracy rate of Base CNN

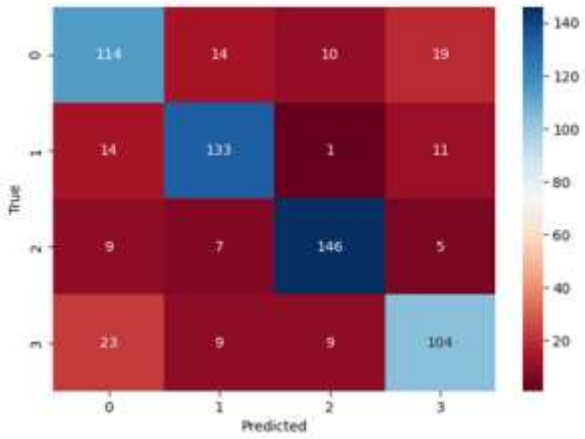

Fig. 7. Confusion Matrix of the Base CNN

TABLE I
RESULT OF QUALITY BASED ON MODELS

Model	Epoch	Train Accuracy (%)	Test Accuracy (%)	Class	Precision (%)	F1-score (%)	Recall (%)
Base CNN	100	92.60	78.98	Cone Cut	71	72	73
				Normal	82	83	84
				Scratch	88	88	87
				Tilted	75	73	72
AlexNet	10 (early stopping)	24.38	24.84	Cone Cut	24	39	100
				Normal	0	0	0
				Scratch	0	0	0
				Tilted	0	0	0
RestNet-50	53 (Early stopping)	92.89	78.03	Cone Cut	74	74	75
				Normal	81	81	81
				Scratch	83	80	77
				Tilted	72	74	77
ViTs	30	75.40	81.34	Cone Cut	75	82	78
				Normal	74	85	79
				Scratch	94	88	91
				Tilted	85	71	78

B. AlexNet

The result in figure 8 shows that the loss rate goes down aggressively during 0 to 2 epochs but after that, it starts to sustain up to 8 epochs while the validation loss went down from 1.7 to 1.3 until the 8 epoch.

https://doi.org/10.31436/ijpcc.v11i2.558

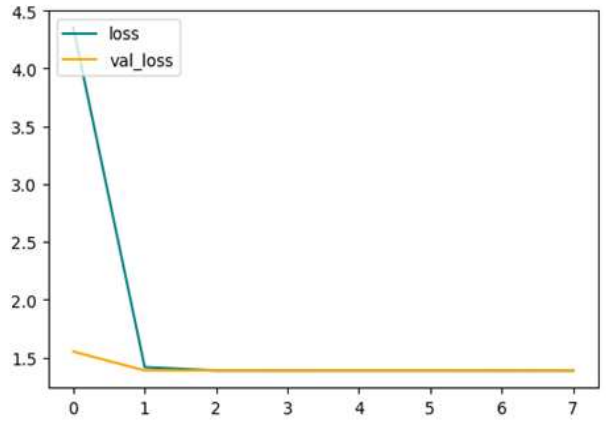

Fig. 8 Figure show the loss rate of AlexNet

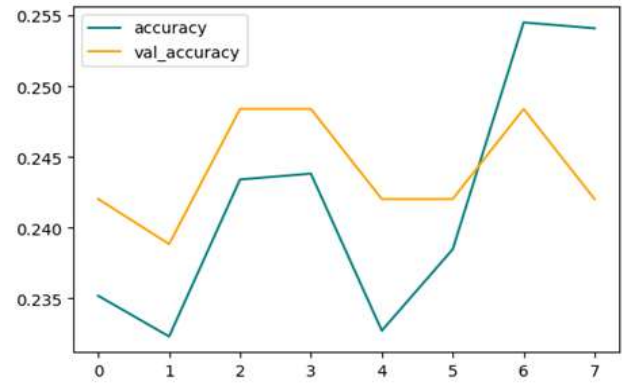

Fig. 9 The accuracy rate of AlexNet

Based on the result in Figure 9, the accuracy of AlexNet is not stabilized throughout the training cycle. The measurement attained after 8 epochs is at most 0.25. Similarly, the validation accuracy goes down after the 6 epochs.

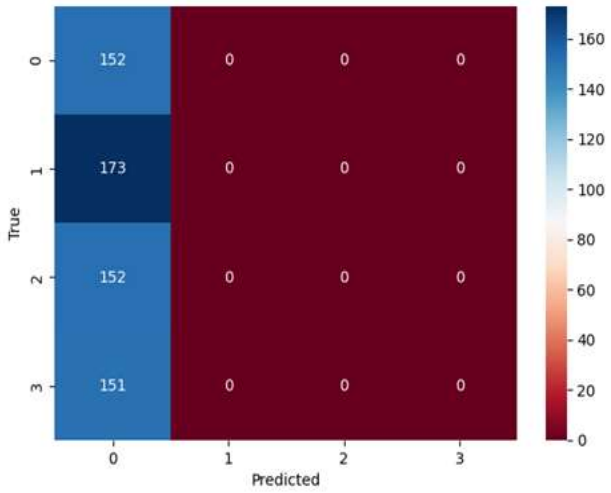

Fig. 10 Confusion Matrix of AlexNet

Based on the result in Figure 10, it can be clearly seen that the 3 classes are missing from being included in the fitting model which is the reason for the low accuracy. Consequently, there is class imbalance that made the cone cut dominate the model prediction with 152 labels as true positive.

C. RestNet-50

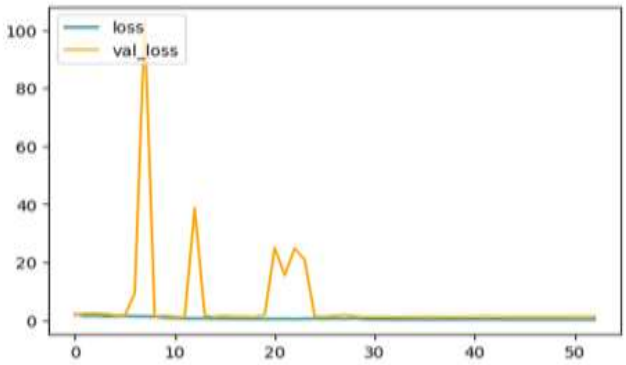

Fig. 11 The Loss Rate of ResNet-50 from 0-53 Epochs.

Based on Figure 11 and Figure 12, the lowest loss rate is 0.14 achieved after 53 epochs during training while the lowest for the validation was achieved after 33 epochs.

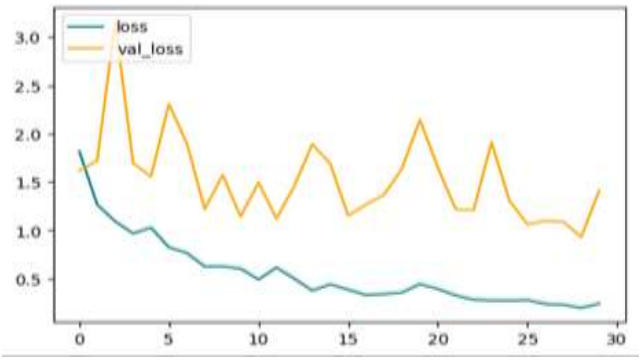

Fig 12 The Loss Rate of ResNet-50 from 0-33 epochs.

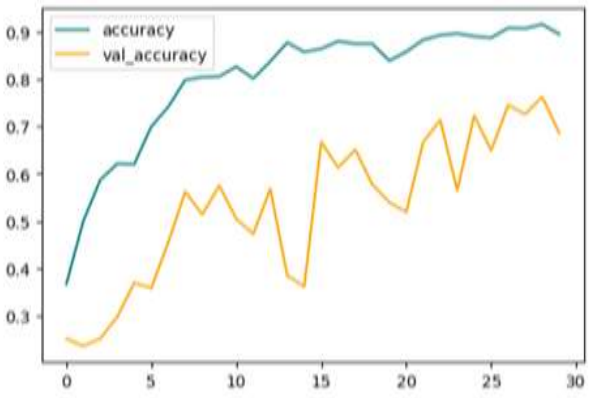

Fig13. The Accuracy Rate of ResNet-50 from 0-53 Epochs.

https://doi.org/10.31436/ijpcc.v11i2.558

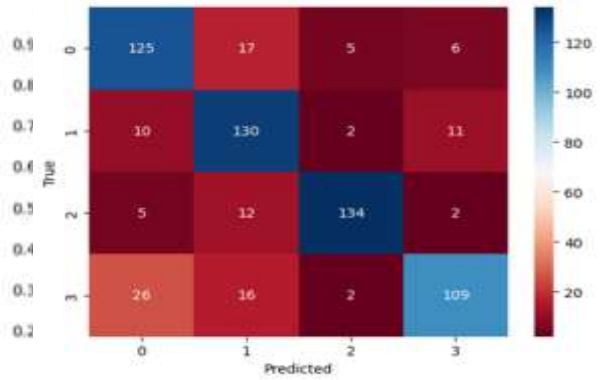

Fig. 14. The Accuracy Rate of ResNet-50 from 0-33 Epochs.

Based on Figure 13 and Figure 14, the accuracy increased gradually after the 7 epochs of training with the highest peak value of 0.92 while the validation rate steadily goes up as the validation cycle goes on to 53 epochs with value of around 0.78.

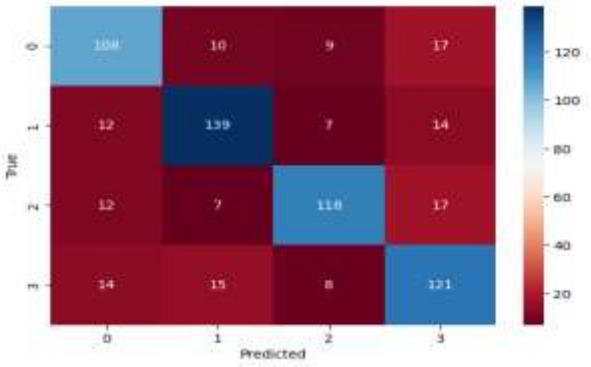

Fig. 15 The Confusion Matrix of ResNet-50.

Based on Figure 15, it can be seen that the number of true positive according to the classes are 108, 139, 118 and 121 respectively, given a total overall accuracy of 486 out of 628 images.

D. VITs

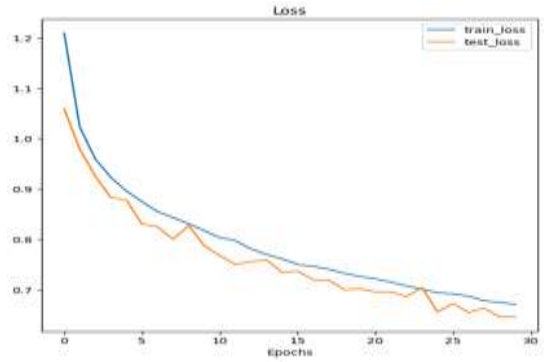

Fig. 16 The Loss Rate of ViTs.

Based on Figure 16, the loss rate during training gradually decreased as the number of epochs increased reaching the highest value of around 0.71 compared to the loss rate on the test set, but it goes to around 0.75.

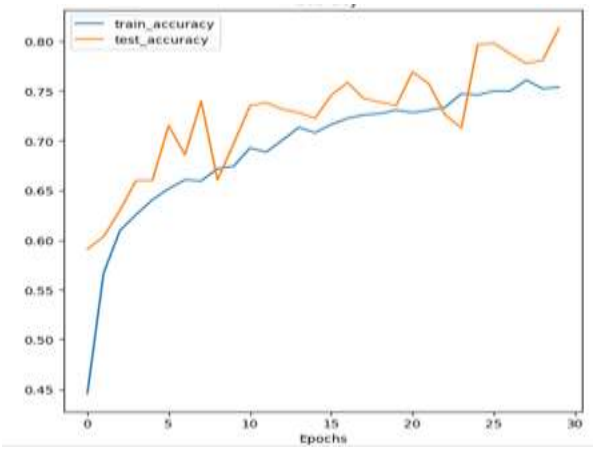

Fig 17. The Accuracy Rate of ViTs.

Based on Figure 17, the training accuracy increased effectively as the number of epochs increasedto the highest value of around 0.75. Meanwhile, the test accuracy starts from 0.59 at zero epoch, and increases around 0.82 at 30 epochs.

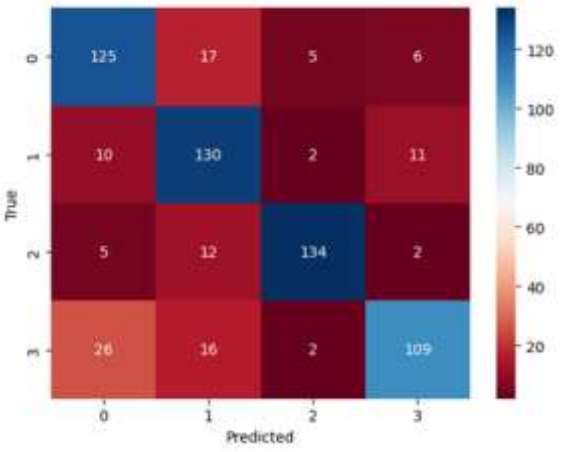

Fig. 18 The Confusion Matrix of ViTs

Based on the confusion matrix in Figure 18, the true positives based on the respective classes are 125, 130, 134 and 109. The total of true positives is 498 out of 612 images. The Scratch has the highest true positive of 134.

V. DISCUSSION

Detection of errors is significantly important because they concern sensitive information in terms of health and privacy of the patients. To produce a panoramic cardiograph with constant image quality common errors must be kept to a

minimum [22]. Common errors, generally include cone cuts, scratch, elongated or deformed, and tilted radiographs. Errors in quality assurance can lead to incorrect diagnoses and inappropriate treatments which can have life-changing impacts on patients. To reduce the possibility of mistakes, it's critical to ensure that the X-ray machine is precisely oriented and that the film or image receptor is well positioned. Even though with time, technology will eventually be able to replace human labor in managing the quality of X-ray images, better quality radiographs may be captured by dental professionals and students with the help of early hardware or software inspection.

This study compared CNN, ResNet-50, AlexNet, and ViTs in terms of how well they could classify radiograph mistakes. This is due to the fact that, based on their abilities, three of the four models that were selected for training produced some encouraging outcomes. Only one model, however, performed poorly when analyzing the images. But because these models rely on specific features and architectural elements, they have unique characteristics when it comes to analyzing X-ray images.

We may infer from the graphs that each model's results are influenced by the specific architectural flow. When using ViTs, the method matched well to keep the train and test cycle close. Assessing the distance between the train and test model for these CNN models shows poor performance was due to models having trouble balancing the training and test accuracy distance. However, ViTs are models that require certain hardware requirements to analyze larger datasets; in other words, they are high-speculation models intended for use with large datasets. Concentious care is needed to incorporate the X-ray images according to their design using the appropriate hyperparameters and data split to optimize the image quality.

Future research can build upon these models to make them even better. The architecture's layers are highly effective in maximizing image accuracy throughout the experiment. It should be noted that the more hidden layers generated, the higher the trainable parameters must be, and the higher the hardware specifications to achieve the requirements. Additionally, a balanced dataset might be a crucial component in incorporating precision to diagnose any class discrepancies between them. Equally, experimenting with splitting data is advised to reduce overfitting or underfitting. In addition, the number of training cycles may additionally have an impact in tracking the peak accuracy of the fitting models.

VI. Conclusion

Radiograph processing is still a laborious procedure. It is essential to create quality assurance solutions that increase practitioners' productivity as well as safeguard patients from false positives and false negatives. In response to this problem, this research proposed various models to detect inaccuracies in bitewing radiographs, experimenting with each model's performance to achieve the best possible performance in terms of accuracy and loss rate. Even though more work is needed to ensure that the CNNs or ViTs were completely satisfactory, the findings could constitute a major advancement in the field. The capabilities of dental specialists may be improved by further development of CNNs and the ViTs paradigm. ViTs-enabled software may be employed due to its flexible training, despite its higher cost. This study demonstrates the value of artificial intelligence and machine learning can potentially provide in medical imaging quality assurance processes.

Acknowledgment

The authors hereby acknowledge the review support offered by the IJPCC reviewers who took their time to study the manuscript and find it acceptable for publishing.

Conflict of Interest

The authors declare that there is no conflict of interest

References

[1] M. A. Barayan et al., "Effectiveness of Machine Learning in Assessing the Diagnostic Quality of Bitewing Radiographs," Applied Sciences (Switzerland), vol. 12, no. 19, Oct. 2022, doi: 10.3390/app12199588.

[2] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, no. 4. 2018. doi: 10.1007/s13244-018-0639-9.

[3] Y. Tian, "Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3006097.

[4] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," Evolutionary Intelligence, vol. 15, no. 1. 2022. doi: 10.1007/s12065-020-00540-3.

[5] K. Han et al., "A Survey on Vision Transformer," IEEE Trans Pattern Anal Mach Intell, vol. 45, no. 1, 2023, doi: 10.1109/TPAMI.2022.3152247.

[6] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, and N. Al Ajlan, "Vision transformers for remote sensing image classification," Remote Sens (Basel), vol. 13, no. 3, 2021, doi: 10.3390/rs13030516.

[7] I. S. Samanta et al., "A Comprehensive Review of Deep-Learning Applications to Power Quality Analysis," Energies 2023, Vol. 16, Page 4406, vol. 16, no. 11, p. 4406, May 2023, doi: 10.3390/EN16114406.

[8] T. Ekert et al., "Deep Learning for the Radiographic Detection of Apical Lesions," J Endod, vol. 45, no. 7, pp. 917-922.e5, Jul. 2019, doi: 10.1016/J.JOEN.2019.03.016.

[9] A. Heidari, S. Toumaj, N. J. Navimipour, and M. Unal, "A privacy-aware method for COVID-19 detection in chest CT images using lightweight deep conventional neural network and blockchain," Comput Biol Med, vol. 145, p. 105461, Jun. 2022, doi: 10.1016/J.COMPBIOMED.2022.105461.

[10] "Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network | upGrad blog." Accessed: Dec. 18, 2023. [Online]. Available: https://www.upgrad.com/blog/basic-cnn-architecture/

[11] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Evolving Deep Convolutional Neural Networks for Image Classification," IEEE Transactions on Evolutionary Computation, vol. 24, no. 2, 2020, doi: 10.1109/TEVC.2019.2916183.

[12] M. Momeny, A. M. Latif, M. Agha Sarram, R. Sheikhpour, and Y. D. Zhang, "A noise robust convolutional neural network for image classification," Results in Engineering, vol. 10, 2021, doi: 10.1016/j.rineng.2021.100225.

[13] M. F. Ibrahim, S. Khairunniza-Bejo, M. Hanafi, M. Jahari, F. S. Ahmad Saad, and M. A. Mhd Bookeri, "Deep CNN-Based Planthopper Classification Using a High-Density Image Dataset," Agriculture 2023, Vol. 13, Page 1155, vol. 13, no. 6, p. 1155, May 2023, doi: 10.3390/AGRICULTURE13061155.

[14] A. O. Tarasenko, Y. V. Yakimov, and V. N. Soloviev, "Convolutional neural networks for image classification," in CEUR Workshop Proceedings, 2019.

[15] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," Neural Computation, vol. 29, no. 9. 2017. doi: 10.1162/NECO_a_00990.

[16] N. A. Mohammed, M. H. Abed, and A. T. Albu-Salih, "Convolutional neural network for color images classification," Bulletin of Electrical Engineering and Informatics, vol. 11, no. 3, 2022, doi: 10.11591/eei.v11i3.3730.

[17] M. M. Krishna, M. Neelima, M. Harshali, and M. V. G. Rao, "Image classification using Deep learning," International Journal of Engineering and Technology(UAE), vol. 7, 2018, doi: 10.14419/ijet.v7i2.7.10892.

[18] A. Ramalingam, "How to Pick the Optimal Image Size for Training Convolution Neural Network? | by Aravind Ramalingam | Analytics Vidhya | Medium," June 24 2021. Accessed: Dec. 18, 2023. [Online]. Available: https://medium.com/analytics-vidhya/how-to-pick-the-optimal-image-size-for-training-convolution-neural-network-65702b880f05

[19] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?," in Advances in Neural Information Processing Systems, 2018.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in 32nd International Conference on Machine Learning, ICML 2015, 2015.

[21] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of image classification algorithms based on convolutional neural networks," Remote Sensing, vol. 13, no. 22. 2021. doi: 10.3390/rs13224712.

[22] Sanjeet Singh, Inderpreet Singh, Farooq Ahmed, and Arshid Baba, "Retrospective Study: Evaluating the Positioning Errors in Digital Panoramic Radiographs," Indian Journal of Contemporary Dentistry, vol. 10, no. 2, 2022, doi: 10.37506/ijocd.v10i2.18413.