PROSTATE CANCER GRADING USING INCEPTION-RESNET

WANG RUI

AFZAN ADAM

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

ABSTRACT

Prostate cancer (PCa) is a common cancer that affects men's health worldwide, and early diagnosis is crucial to improve patient survival. This study aims to improve the efficiency and reliability of prostate cancer diagnosis through pathological image analysis and artificial intelligence technology. The Gleason grading system has been a key tool for assessing the aggressiveness of prostate cancer. Existing computer science research has explored various prostate cancer classification algorithms, revealing successes and limitations. Despite significant progress, optimal performance and practical applicability remain elusive. This project uses two main datasets: the Prostate Cancer Grading Assessment (PANDA) dataset and the Hospital UKM (HUKM) dataset. The PANDA dataset contains 11,000 whole slide images (WSIs), while the HUKM dataset contains 193 regions of interest (ROIs) extracted from WSIs. Two pre-trained models, Inception-ResNet and EfficientNet, are used for transfer learning training on the PANDA dataset, respectively, and the trained PANDA model is directly applied to the HUKM dataset for testing. The final EfficientNet model achieved 61% accuracy, and after transfer learning and optimization for the HUKM dataset, it finally achieved 90% accuracy on the test set. In addition, the project also used Streamlit to develop a user interface that enables users (especially clinicians) to upload pathology images and quickly obtain accurate results. This helps to make diagnoses faster and more accurately, ultimately saving medical professionals valuable time.

Key word: Prostate cancer; Gleason score; PANDA; Inception-Resnet; EfficientNet; Finetuning

INTRODUCTION

Prostate cancer (PCa) is the second most diagnosed cancer in men worldwide, accounting for 7.3% of the most commonly diagnosed cancers, and is one of the fifth leading causes of cancer death. The five-year survival rate for PCa patients diagnosed early is over 90%, whereas if it is detected only in the late or highly metastatic stage, the survival rate is only 30% (Leslie 2023). Prostate cancer is a cancer that occurs in the prostate, a small walnut-shaped gland in men that produces semen that nourishes and transports sperm. Usually, prostate cancer grows slowly, is initially localized to the prostate, and may not cause serious damage. However, while some types of prostate cancer grow slowly and may require little or no treatment, other

types are aggressive and can spread quickly (Mayo Clinic 2022). The malignancy of prostate cancer can be evaluated by histological grading. The most used is the Gleason scoring system. The malignancy of prostate cancer is classified into grades 1 - 5 based on the sum of the scores of the main structural areas and secondary structural areas in the prostate cancer tissue.

Developing robustness in predictive models is a challenging task. Images of the same disease may vary greatly due to differences in acquisition parameters, equipment, time, and other factors. This results in poor robustness and generalization of existing deep learning models. Therefore, improving the model structure and training methods by existing technology to improve the generalization ability of deep learning is one of the key directions in the future (Wang 2022). Furthermore, in recent years, deep learning technology has achieved tremendous popularity in prostate cancer detection and Gleason grading with the help of massively parallel computing (GPU) (Linkon 2021).

The goal of this project is to use histopathology images for accurate detection and grading of prostate cancer to meet clinical needs. Currently, this process relies primarily on manual analysis by pathologists, which results in subjective interpretation, as well as inter-observer variability and time-consuming workflows. To address these issues, this project aims to leverage existing deep learning models to handle Gleason scoring.

In this project, the initial phase involves leveraging the Inception-ResNet algorithm for training and optimization on the publicly available PANDA dataset. Advanced techniques including feature extraction and transfer learning of pathological images will be used to enhance the performance of the model. Subsequent evaluation of model performance will be performed on the local HUKM dataset. This project aims to introduce flexibility and extend its applicability through the inclusion of the PANDA Challenge dataset, thereby facilitating a comprehensive evaluation and comparison of the model's performance with other convolutional neural network architectures. This comprehensive approach aims to provide clinicians with a more accurate and efficient tool to detect and classify prostate cancer grade.

METHODOLOGY

This project uses the incremental model, a software development method that can be used to build large and complex systems. It is based on the idea of adding new features or "increments" to an existing system instead of building the entire system from scratch at once. Improvements can be made gradually, models can be implemented slowly, and different functional modules can be handled independently. It is very important to achieve complexity and maintainability in the prostate cancer grading model. Prostate cancer grading is a complex task, and the incremental model allows large problems to be broken down into smaller, more manageable parts, and as the project progresses, we can adjust subsequent development plans based on intermediate results(See figure 1).

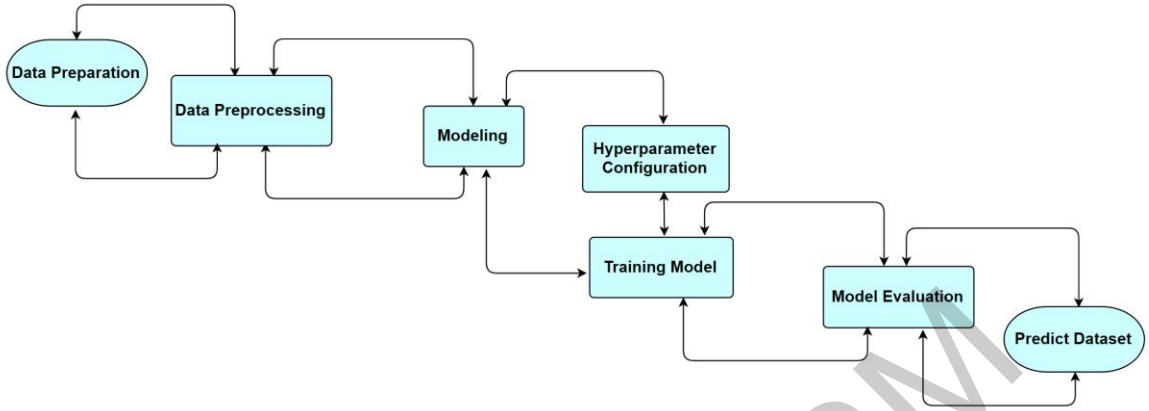

Figure 1 Incremental Model

This study is divided into three main parts. The first part is training and evaluation on the PANDA dataset. The second part is to improve the trained model and test it on the HUKM dataset. The third component provides access to the model through a web-based system. This will provide medical professionals with the tools to quickly and accurately diagnose prostate cancer, thereby improving treatment outcomes and patient survival rates.

1. Dataset

This study mainly uses two datasets: PANDA (Prostate Cancer Grading Assessment) dataset and HUKM (Hospital UKM) dataset.

The PANDA dataset contains about 11,000 whole slide images (WSI), which provides large-scale training data for this study. These images are stored in TIF format to ensure high-quality image detail preservation. In the PANDA dataset, the number of samples for each Gleason score category is about 1,300.

The HUKM dataset contains 193 regions of interest (ROIs), which are extracted from WSI. The images of the dataset are also stored in TIF format. The category distribution of the HUKM dataset is as follows:

- Benign: 52 samples
- Gleason grade 3 (G3): 41 samples
- Gleason grade 4 (G4): 56 samples
- Gleason grade 5 (G5): 44 samples

2. Pre-processing

In prostate cancer pathology images, some areas may appear blank or overstained. These areas have no practical significance for the diagnosis and grading of cancer. Instead, they increase the computational burden and affect the performance of the model. Therefore, these useless images are filtered out before training the model. Specific implementation: Split a large whole slice image (WSI) into small blocks of 256x256 pixels to prepare for subsequent analysis and

classification. By visualizing the segmented small images, the details and organizational structure of the image can be intuitively observed to ensure the accuracy of the segmentation process(See Figure 2).

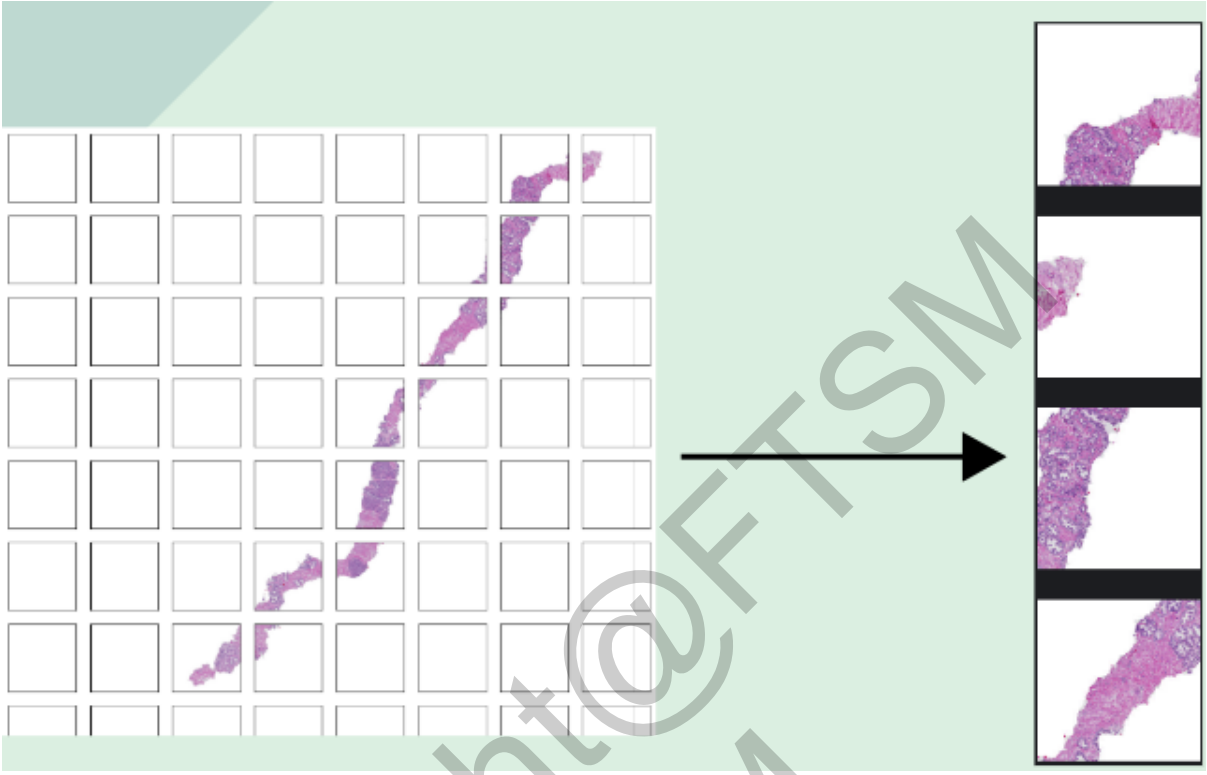

Figure 2 Pre-processing dataset

3. Model

Figure 3 shows the InceptionResNetV2 architecture combines the inception and residual connection models to improve performance. This hybrid approach enables the network to take advantage of the strengths of both models, including faster training time and avoiding the vanishing gradient problem. Residual connections also allow the network to skip some layers during training. In addition, InceptionResNetV2 uses multiple kernel sizes in a single layer to extract patterns with different hierarchical structures, further enhancing the network's ability to capture features of different complexities (Mehdi Neshat et al. 2024).

Figure 4 shows architecture of EfficientNetB0 with MBConv as basic building blocks. EfficientNet uses a technique called compound coefficient to scale up models in a simple but effective manner. Instead of randomly scaling up width, depth, or resolution, compound scaling uniformly scales each dimension with a certain fixed set of scaling coefficients. Using this scaling method and AutoML, the authors of EfficientNet developed seven models of various dimensions, which surpassed the state-of-the-art accuracy of most convolutional neural networks, and with much better efficiency (GeeksforGeeks, 2024).

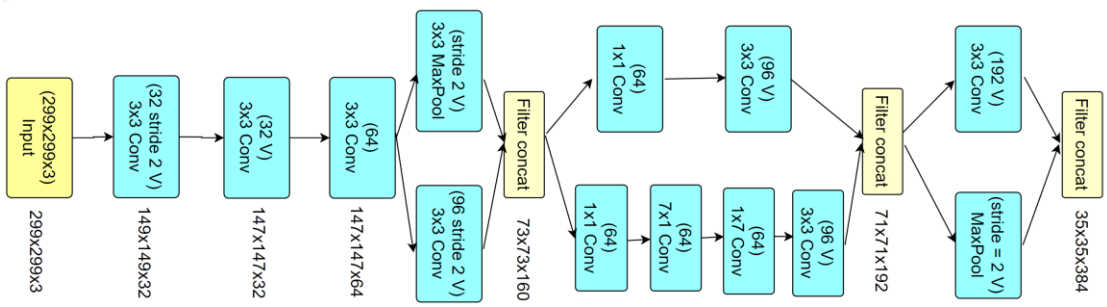

Figure 3 Inception ResNet V2 stem

Source: GeeksforGeeks 2022

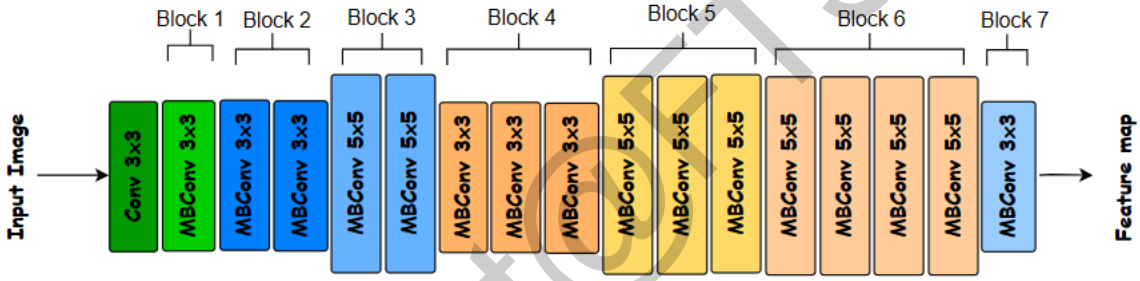

Figure 4 Architecture of EfficientNetB0

Source: Ahmed & Sabab, 2020

4. Fine -Tuning and Transfer Learning

Fine-tuning in deep learning is a form of transfer learning. It involves taking a pre-trained model, which has been trained on a large dataset for a general task such as image recognition or natural language understanding, and making minor adjustments to its internal parameters. The goal is to optimize the model's performance on a new, related task without starting the training process from scratch (Amanatullah, 2023).

Transfer learning is an approach to machine learning where a model trained on one task is used as the starting point for a model on a new task. This is done by transferring the knowledge that the first model has learned about the features of the data to the second model. In deep learning, transfer learning is often used to solve problems with limited data. This is because deep learning models typically require a large amount of data to train, which can be difficult or expensive to obtain (Fagbuyiro, 2024).

5. Model Evaluation

Model evaluation is a crucial step in the development process of machine learning models. In order to comprehensively measure the performance of the model, this project will use the following evaluation criteria:

- Accuracy: The proportion of the number of samples correctly classified by the model to the total number of samples.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

- Precision & Recall: Precision is the proportion of cases where the model predicts a positive class that actually is a positive class. And the recall rate is the proportion of cases where the model correctly predicts a positive category when it is a positive category.

$$\text{Precison} = \frac{True\ positive}{True\ positive + False\ positive}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}$$

- F1-Score: The harmonic average of precision and recall, used to consider both.

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

- Confusion Matrix: Shows the combination of true labels and predicted labels, used to analyze the classification performance of the model.

6. Model deployment and system implementation

Deploy the trained model to a Streamlit-based Web system, providing a friendly user interface and a comprehensive interactive experience. The specific steps are as follows:

1. Export and deploy the trained model.
2. Use Streamlit to build a user interface, providing user login, file upload, result viewing, and history functions.

- Implement user authentication to ensure that only legitimate users can access the system.
- After the user uploads the pathological image, the system calls the model to predict and display the analysis results.
- The user views the prediction results and can access the history of previous tests.
- After the user completes the operation, he can log out safely through the system's logout function to ensure data security

RESULT AND DISCUSSION

1. Developing the pre-trained model

After training the Inception-ResNet pre-trained model on the PANDA dataset, the results show that the training and validation accuracies are both around 40%. During the training process, ReduceLROnPlateau was used to dynamically adjust the learning rate, EarlyStopping was used to prevent overfitting, Keras Tuner was used to optimize hyperparameters, and RandomSearch was used to search for hyperparameters, but the effect was still not ideal.

The main reason is that due to the limitation of GPU resources, it is impossible to perform multiple rounds (epochs) of training on Kaggle, which limits the training time and training effect of the model and makes it difficult for the model to fully learn the data features. It is speculated that InceptionResNet-V2 may be too complex for this task and difficult to converge within the limited training time. In addition, the characteristics of the PANDA

dataset may be very different from the pre-training dataset of InceptionResNet-V2, resulting in poor transfer learning of the model. Therefore, the EfficientNet B0 model was used instead, which is a lightweight model design with excellent performance and efficiency under resource-constrained conditions.

In contrast, after training the entire dataset for 20 epochs using the EfficientNet pre-trained model, the model performance improved, with an accuracy of 61% and the QWK (quadratic weighted Kappa) of 0.83. Figure 5 shows the detailed results of the training process.

By comparison, it can be found that EfficientNet performs better than Inception-ResNet on the PANDA dataset. This may be because the architecture of EfficientNet is more suitable for the characteristics of the dataset, with relatively low complexity, and is easier to converge within a limited training time.

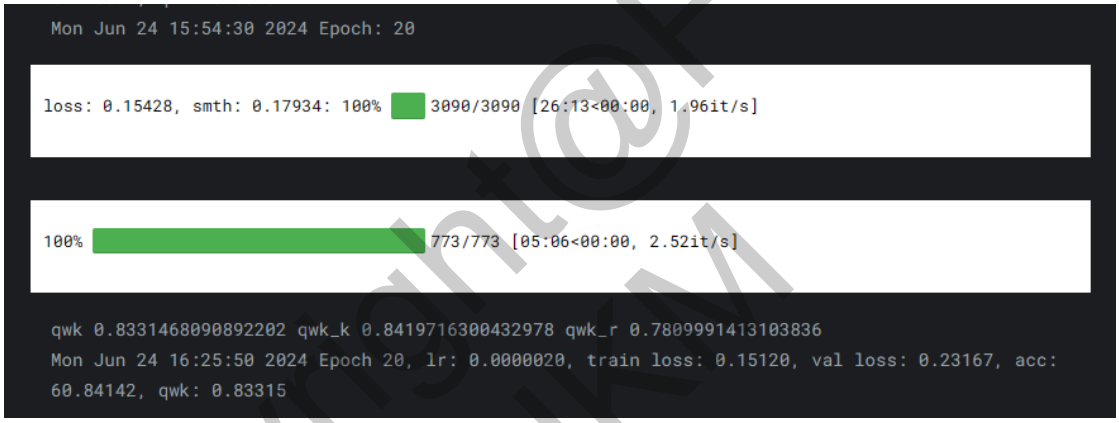

Figure 5 The result of EfficientNet

2. Test on HUKM dataset

After testing on the HUKM dataset, the classification result is 40%. Looking at the confusion matrix in Figure 6, it is found that the model has certain misclassification phenomena in the classification of benign and malignant images at all levels, especially on G3. To summarize the reasons for the unsatisfactory test results, it is first found that during the PANDA training process, the data labels are not well mapped to the HUKM dataset, resulting in misjudgment in the prediction. Secondly, there are some differences between the two datasets in terms of data glandular morphology, blood vessels, glandular structure, nuclear characteristics, matrix, etc., which will introduce additional variability in the classification.

	precision	recall	f1-score	support
Benign	0.58	0.35	0.43	52
G3	0.24	0.59	0.34	41
G4	0.56	0.27	0.36	56
G5	0.63	0.50	0.56	44
accuracy			0.41	193
macro avg	0.50	0.42	0.42	193
weighted avg	0.51	0.41	0.42	193

Figure 6 Confusion index tested in HUKM

3. Train the pre-trained model and test in HUKM dataset

According to the loss curve in Figure 7, the training loss (blue line) and validation loss (orange line) show an overall downward trend, indicating that the model gradually improves during the learning process. The loss drops rapidly in the early stage of training (around 0-5 epochs), and then the rate of decline slows down. The validation loss fluctuates greatly in the first few epochs, and then tends to decrease steadily. The gap between the training loss and the validation loss is not large, and there is no obvious overfitting phenomenon.

According to the accuracy curve in Figure 8, the training accuracy (blue line) and validation accuracy (orange line) show an overall upward trend. The validation accuracy fluctuates greatly in the first 30 epochs, and then stabilizes between 80% and 90%. The training accuracy rises relatively smoothly, eventually reaching about 85%. The validation accuracy suddenly rises in some epochs (such as between 25-30 epochs), which may be due to fluctuations caused by a small data set.

Figure 7 Training and validation loss curve

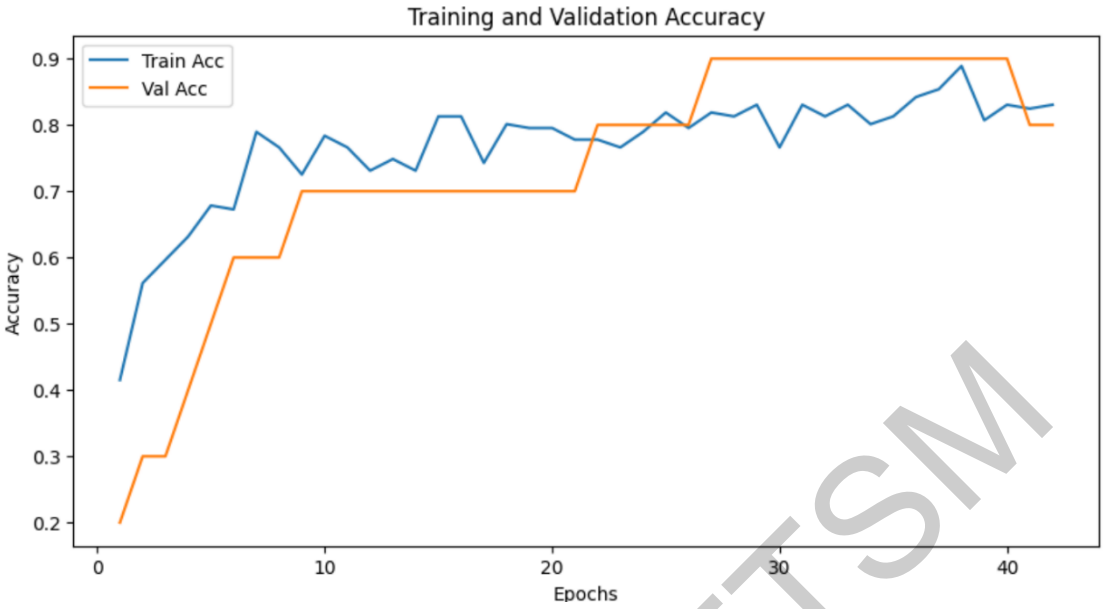

Figure 8 Training and validation accuracy curve

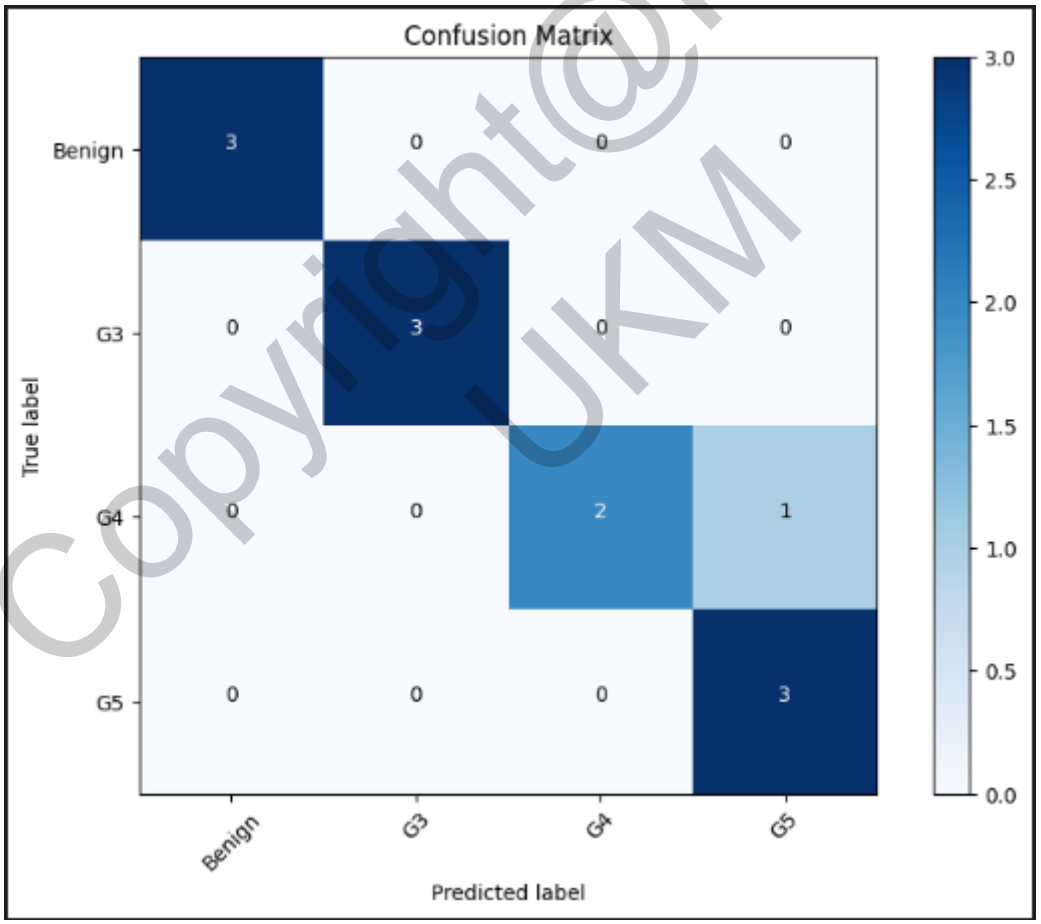

Figure 9 Confusion metric after pre-trained model

The final training results: On the training set, the training loss is 0.48 and the accuracy is 85.38%. On the validation set, the validation loss is 0.37 and the accuracy is 90%. On the test set, the test loss is 0.51, which is slightly higher than the training and validation losses, but still within an acceptable range, and the test accuracy is 91.67%. The precision is 0.93, the recall is 0.91, and the F1 score is 0.91. These indicators are all very high, close to 1, which generally shows that the model performs well in all categories.

Table 1 Evaluation Result

Test Accuracy	0.90
Validation accuracy	0.91
Precision	0.93
Recall	0.91
F1 Score	0.91

4. Deploy the final trained model to the system.

The interface includes a user authentication system that ensures secure access to the application (See figure 10). The main features include:

User Registration

Users can create a new account by providing the necessary details.

Login

Registered users can log in to the application.

Password Change

Users can change their password securely.

Logout

Users can log out of the application and end the session.

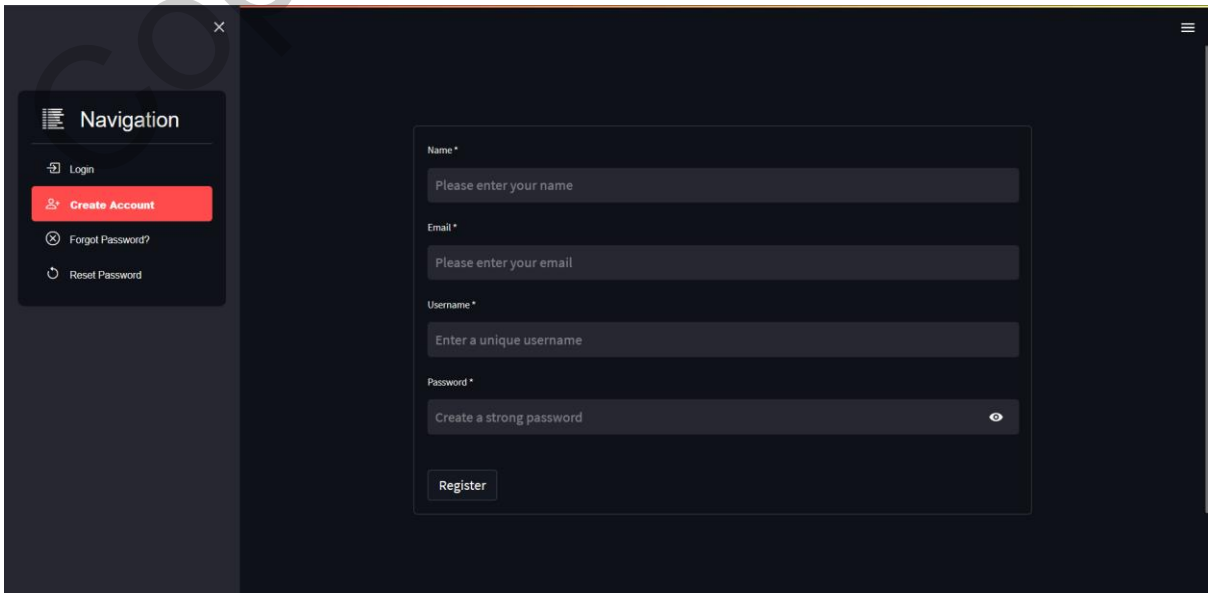

Figure 10 User registration

After being authenticated, users can upload prostate cancer images for classification. The system processes the image and determines the category it belongs to, and then provides the classification result to the user.

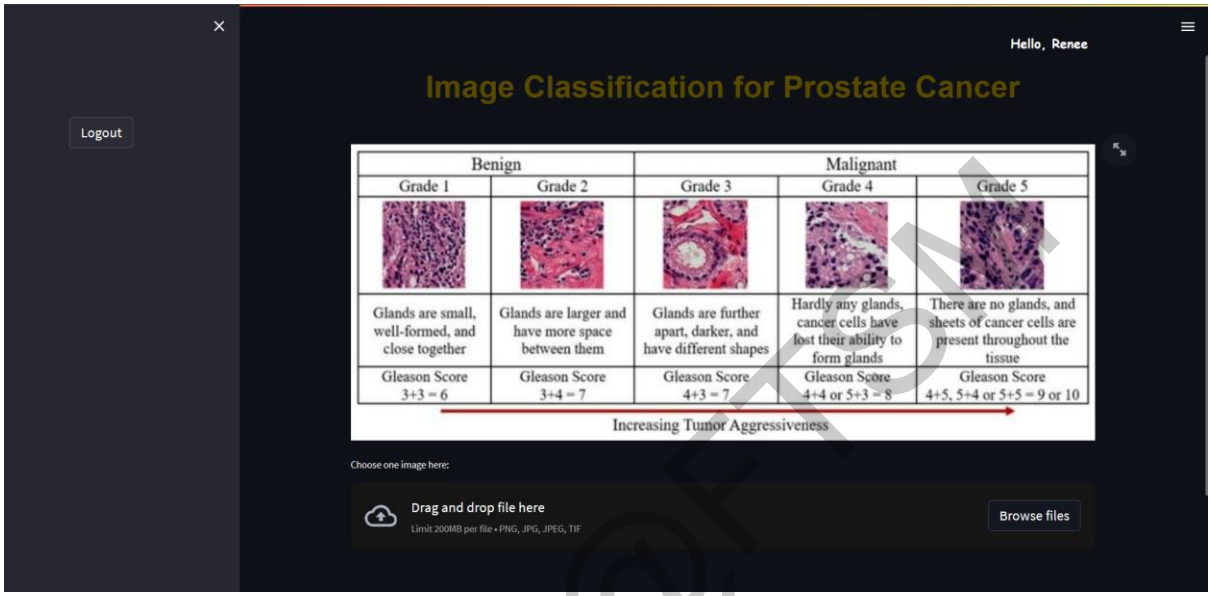

Figure 11 System main page

Figure 12 is the result page after the user uploads the image, which shows the classification results and detailed confidence scores and provides feedback options.

Figure 12 Image classification results

CONCLUSIONS

The project aims to improve the efficiency and reliability of prostate cancer diagnosis through pathological image analysis and artificial intelligence technology. The project uses two pre-trained models: Inception-ResNet and EfficientNet. After transfer learning on the PANDA dataset, the EfficientNet model performed better with an accuracy of 61%. The model was then applied to the HUKM dataset and further optimized, achieving 90% accuracy on the test set. In addition, a user interface was developed using Streamlit to enable clinicians to upload pathological images and quickly obtain accurate diagnosis results. The system is expected to speed up diagnosis and improve accuracy, thus saving valuable time for medical professionals.

Although the differences between the datasets posed some challenges, these obstacles were successfully overcome through transfer learning and model optimization. The main innovation of the project is to combine a large-scale public dataset with the UKM hospital dataset to improve the flexibility of the model and the diversity of the dataset.

In the future, it is necessary to further improve the generalization ability of the model, adjust the model parameters, and explore deep learning architectures..

ACKNOWLEDGEMENT

I am very fortunate to have Dr. Afzan Adam as my supervisor. Thank you for the guidance and support you have provided. I am grateful to everyone who has contributed to this project. Your assistance has been invaluable.

REFERENCE

Neshat, M., Ahmed, M., Askari, H., Thilakaratne, M., & Mirjalili, S. (2024). Hybrid Inception Architecture with Residual Connection: Fine-tuned Inception-ResNet Deep Learning Model for Lung Inflammation Diagnosis from Chest Radiographs. Procedia Computer Science, 235, 1841–1850. https://doi.org/10.1016/j.procs.2024.04.175.

Prostate cancer - Symptoms and causes - Mayo Clinic. (2022, December 14). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087.

Rodler, S., Kopliku, R., Ulrich, D., Kaltenhauser, A., Casuscelli, J., Eismann, L., Waidelich, R., Büchner, A., Butz, A., Cacciamani, G., Stief, C. G., & Westhofen, T. (2023). Patients' Trust in Artificial Intelligence–based Decision-making for Localized Prostate Cancer: Results from a Prospective Trial. European Urology Focus. https://doi.org/10.1016/j.euf.2023.10.020

Wang, L. (2022). Deep learning techniques to diagnose lung cancer. Cancers, 14(22), 5569.

Linkon, A. H. M., Labib, M. M., Hasan, T., Hossain, M. T., & Jannat, M. (2021). Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. Informatics in Medicine Unlocked, 24, 100582. https://doi.org/10.1016/j.imu.2021.100582.

GeeksforGeeks, "Lung Cancer Detection using Convolutional Neural Network CNN," GeeksforGeeks, Sep. 06, 2022. https://www.geeksforgeeks.org/lung-cancer-detection-using-convolutional-neural-network-cnn/.

Ahmed, T., & Sabab, N. H. N. (2020, September 27). Classification and understanding of cloud structures via satellite images with EfficientUNet. arXiv.org. https://arxiv.org/abs/2009.12931.

GeeksforGeeks. (2024, June 3). EfficientNet Architecture. GeeksforGeeks. https://www.geeksforgeeks.org/efficientnet-architecture/

Fagbuyiro, D. (2024, April 20). Guide to Transfer Learning in Deep Learning - David Fagbuyiro - Medium. Medium. https://medium.com/@davidfagb/guide-to-transfer-learning-in-deep-learning-1f685db1fc94

Amanatullah. (2023, September 21). Fine-Tuning the Model: What, why, and how - Amanatullah - medium. Medium. https://medium.com/@amanatulla1606/fine-tuning-the-model-what-why-and-how-e7fa52bc8ddf

Wang Rui (A184975)
Ts.Dr.Afzan Adam
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia