



Name – Shumaila Naaz

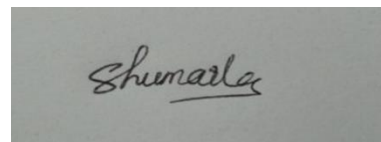
MCA 3rd Semester July-23

Enrolment No – A9929723001859(EI)

**BREAST HISTOPATHOLOGY IMAGES USING DEEP
LEARNING ALGORITHM IN PYTHON**

DECLARATION

I, student of Master of Computer Application hereby declares that the Breast Histopathology Images Using Deep Learning Algorithm in Python, which is submitted by me to, Amity University Online, Noida, Uttar Pradesh in partial fulfilment of requirement for the award of the degree of Master of Computer Application, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

A rectangular box containing a handwritten signature in black ink. The signature appears to be 'Shumaila' written in a cursive style.

Name and signature of student

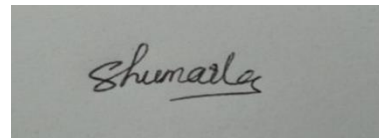
PLAGIARISM REPORT

This is to certify that I of **Master of Computer Application Semester III**

Minor Project/Seminar (Evaluation) Enrolment no. A992972300185 **(EI)**

Session 2023-2024 has submitted the report.

The plagiarism has been checked through tool studymoose.com and it came out to be
75.9% Unique.

A rectangular box containing a handwritten signature in black ink. The signature appears to be 'Shumaila' written in a cursive style.

Name and signature of student

Table of Contents

1. Chapter 1	Introduction	6
2. Chapter 2	Aim and Objectives	6
3. Chapter 3	Literature Review	7
4. Chapter 4	Methodology	11
5. Chapter 5	Data Interpretation.....	14
6. Chapter 6	Conclusion and Recommendation.....	15
7. Chapter 7	Project Timeline	16
8. Chapter 8	References	17

Introduction

Data analysis is going to rely on deep learning algorithms in Python programming language to work on breast histopathology images using an observational secondary data collection approach. The “Breast Histopathology Images” used in this study can be retrieved from “kaggle.com” and contains image format data. This research method is quantitatively grounded and will adopt the deductive approach regarding the positivism paradigm. The process will be done using the help of Google Collab software to enhance computational proficiency and sharing. In order to analyse the dataset effectively, the EDA methods can be applied to the data. After EDA, data normalization and preprocessing measures will be performed to create the desired model. The data will be processed by using the deep learning models CNN and the VGG16 structure to interpret the data more specifically and reach valuable conclusions. These models will be done based on various evaluation parameters with a view of judging the efficiency of the models in the classification and analysis of breast histopathology images, towards improvement of medical image analysis.

Aim and Objectives

Aim

This project aims to analyse breast histopathology images using deep learning algorithms to develop an accurate classification model for medical image analysis.

Objectives

- To use the “Breast Histopathology Images” data set from “Kaggle.com” for analysis by using secondary data collection techniques
- To use quantitative methods and deductive approaches to perform this analysis

- To prepare the data for feeding it to deep learning models and to obtain a more meaningful representation
- To use CNN and VGG16 models due to image classification
- To assess the applied models and qualitative and quantitative measures to check the efficacy of the models effectively

Literature Review

Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models

Breast cancer is still a major public health issue since it is one of the top killers of women suffering from cancer around the world. Getting a diagnosis at a young age can help increase the survival rate of cancer, even as the current methods of diagnosis present specific challenges; biopsy and histopathological analysis. In order to overcome these drawbacks, the major focus has been paid to the development of automated diagnostic systems which can take a paramount significance in improving the diagnostic results and efficacy (Hameed et al. 2020). An ensemble deep-learning model is proposed for the classification of non-carcinoma and carcinoma breast cancer histopathology images using the proposed dataset.

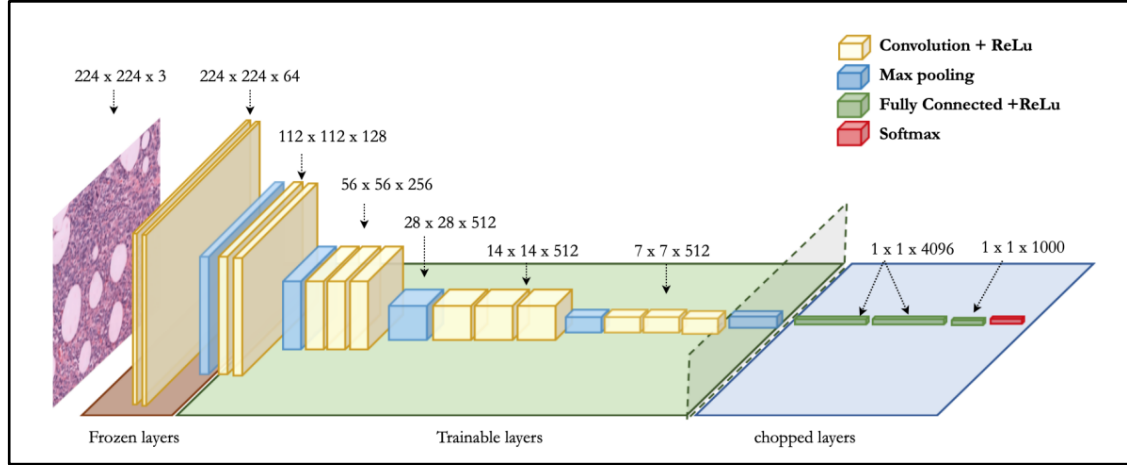


Figure 1: Representation of fine-tuned VGG16 architecture

(Source: Hameed et al. 2020)

Four deep learning models using the VGG16 and VGG19 are used for training and testing with the help of 5-fold cross-validation methods. The proposed combination of fine-tuned VGG16 and VGG19 models provided even better performance, more specifically in identifying carcinoma. The proposed ensemble approach achieved a carcinoma detection sensitivity of 97.73%, an overall accuracy of 95.29% and an F1 Score of 95.29% (Hameed et al. 2020). These results demonstrate the high capability of the proposed approach for automated classification of complex histopathological images, paving the way for potential operative diagnosis in clinical scenarios.

Breast Cancer Classification Using Deep Learning Approaches and Histopathology Image: A Comparison Study

Based on this project, CNN models are one of the most crucial deep-learning architectures used for breast cancer classification. This paper compares and evaluates different deep learning patterns for breast cancer histopathology image binary, four-class, and eight-class image classification. In the models' architectures, several considerations such as pre-processing techniques, data augmentation, and transfer learning affect the model accuracy (Shahidi et al.

2020). Focusing on the relations between these components, the paper underlines their importance in obtaining a higher classification rate in the datasets of histopathological images. In addition, the study finds four modern prototypes, namely ResNeXt, Dual Path Net, SENet, and NASNet, effectively tested on the ImageNet database.

These advanced models are assessed in binary, four-class, and eight-class tasks using the BreakHis and BACH datasets, and the results are compared to previous works. The work uses the InceptionResNet-V2 network which is known to be the best for binary and eight-class classifications and thus serves as the base for the detailed comparison (Shahidi et al. 2020). The given work provides a systematic approach to experimental settings, moreover, it is useful for understanding the current benchmarks in breast cancer histopathology classification.

A new deep convolutional neural network model for classifying breast cancer

histopathological images and the hyperparameter optimisation of the proposed model

The convolutional neural network surprisingly has striking performance in medical diagnosis as well as in the analysis of images, which in turn enhances many fields including but not limited to drug discovery, time series modelling and optimization. In the histopathologic breast cancer image analysis, difficulties come from the fact that images contain significant likenesses and healthy and cancerous tissues both exist in various parts of whole-slide images. These factors make vulnerability coupled with the reliability of detecting and timely diagnosing tumours cardinal to the treatment of breast cancer (Burçak et al. 2021). This is possible to presuppose that the application of deep learning for classifying breast cancer histopathological images is not only beneficial for overcoming the pathologists' workload but also reduces the subjectivity of diagnostics and stabilizes the quality of classification.

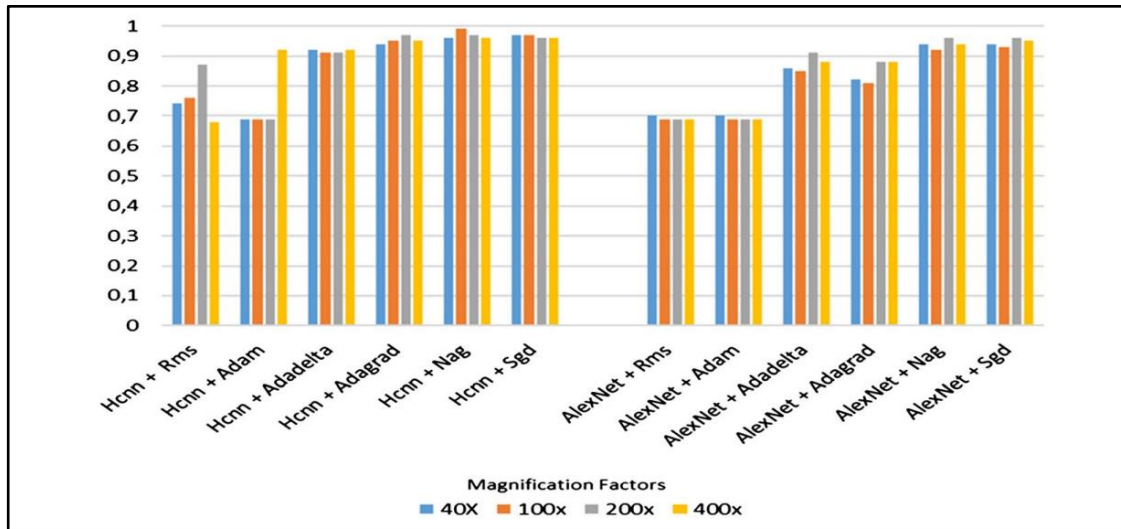


Figure 2: Comparison of model performances

(Source: Burçak et al. 2021)

To overcome these challenges, this work presents a deep CNN model. The model uses optimization algorithms, SChristine, Stochastic gradient descent Nesterov accelerated gradient AdaGrad RMSprop AdaDelta and Adam to set weights and fine-tune for backpropagation. Using parallel computing architecture and with CUDA-enabled GPU, the model gains high training speed with a limited training hardware resource. Qualitative results serve as proof that the proposed model works by reaching a perfect classification rate that is 99.05% (Burçak et al. 2021). These results show that trained deep learning networks with rich and complex architectures can offer accurate and quick breast cancer diagnosis to improve the right decision during treatment.

Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images

Tumour grade is the most prognostic factor in breast cancer patients and though formally assessed visually by pathologists, disagreed largely interobserver-wise. This paper proposes a model for automatic imaging of breast cancer grade using whole slide histopathology images

with no need for manual annotations. The training was performed on 706 young invasive breast cancer patients with age <40 and evaluated tumour grade low-intermediate and high and its parameters like nuclear grade, tubule formation, and mitotic rate (Wetstein et al. 2022).

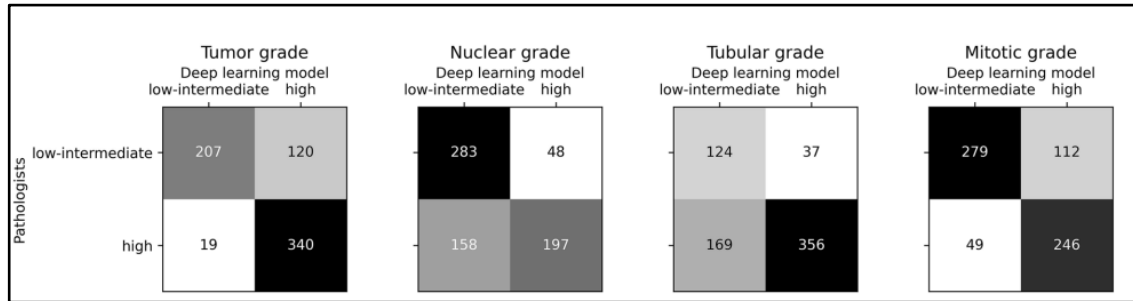


Figure 3: Confusion matrices for no special type (NST) tumour grading and grade components between pathologists and the deep learning model

(Source: Wetstein et al. 2022)

Using the dataset of 686 mutually exclusive test samples with annotations from pathologists as the reference standard, the model provided 80 % accurate results, with a Cohen's Kappa of 0.59. The proposed groups demonstrated statistically significant prognosis of overall survival (OS) and disease/recurrence-free survival (DRFS/RFS) using survival analysis (Wetstein et al. 2022). Observed for hazard ratios, statistical significance remained elusive after multivariable adjustment using clinicopathologic characteristics. The application of this model shows its feasibility in generating repeatable and automated breast cancer grading to clinical factors.

Methodology

The analysis will follow a clear structure in order to categorise breast histopathology images applying deep learning methods. Secondary data sources will be gathered from the "Breast Histopathology Images" source of "Kaggle.com". Google Collab will be entirely used as the computing environment because of its compatibility with deep learning frameworks and the

availability of free GPU (Desai et al. 2021). Use of EDA will be utilised as a method to investigate the data and its visual presentation. The distribution of data will be analysed and further, possible pattern and pattern irregularities will have to be solved. This is because before running any deep learning method, data preprocessing and normalization will be done about the image input.

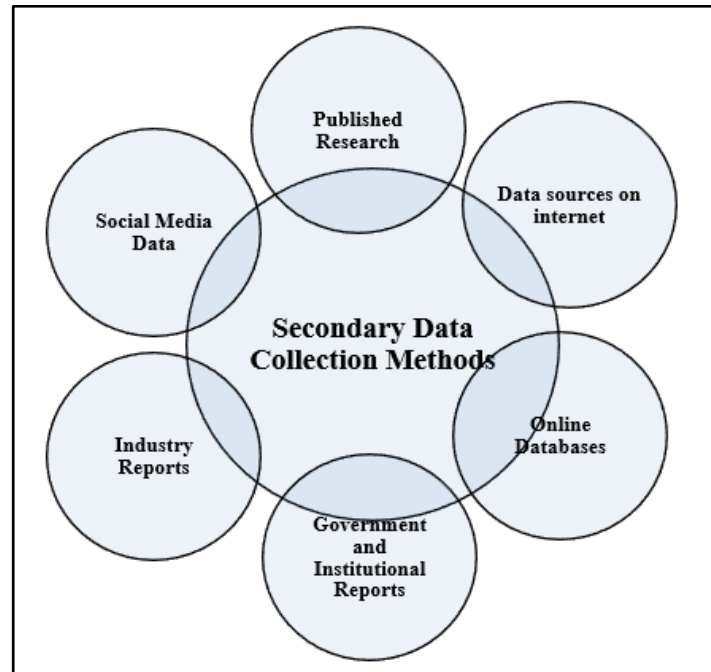


Figure 4: Secondary Data Collection Process

(Source: Desai et al. 2021)

This pre-processed data will then be input into the deep learning models particularly CNN or Convolutional Neural Network and VGG16. The CNN model will therefore act as a reference for this performance benchmark model whereas the VGG16 model will implement transfer learning to increase the accuracy and efficiency levels. Along with this model training, the dataset will be divided into train, validation and test sets to achieve an accurate assessment of its performance (Naveen and Diwan, 2021). In order to fix the model tuning of hyperparameters in the classification models for the best outcome will be done. Once training is over the

performance of the model will be evaluated for accuracy, precision, recall rate, F1 score and confusion matrix.



Figure 5: Process of Quantitative Method

(Source: Naveen and Diwan, 2021)

Ethical issues will be a central focus of the current discussion at all stages of the assessment. This data set will be public and will be obtained from sources that allow only research usage hence passing policy checks. In order to ensure the patient's related data will not compromise the anonymity of the patients, none of the personally identifiable data will be used in the dataset (Desai et al. 2021). The analysis will therefore be conducted with true transparency and the method and process used shall be well documented. The conclusions from the analysis will be presented in a socially responsible manner while respecting the findings' limitations so as not to misrepresent them. Several kinds of knowledge gained or theories discovered will not be used for other purposes, particularly within treatment settings without investigation (Naveen and

Diwan, 2021). It will also prevent vices such as altercations, favouritism, deceit and lots of unnecessary evils which are un-beaconing on research integrity.

Data Interpretation

The analysis of the predicted outputs will be on comparing the performance of the CNN and VGG16 in diagnosing breast histopathology images. Coupled with tuning the models, the objectives are to train and test the models and compare the predicted outputs to evaluate the models that reveal cancerous and non-cancerous cells. According to the five-fold, the overall error rate, accuracy, precision, recall rates, F1-score, and confusion matrix shall be employed since the models' objective is to correctly classify the given images (Desai et al. 2021). The measures of accuracy will describe the share of correct classifications, and, therefore, high accuracy is perhaps the best strategy for classification. Precision and recall evaluate the models for their capacities to detect the true positives and minimize false negatives. The F1 score will help give a measure of both precision and recall, even when there are issues with high-class imbalance.

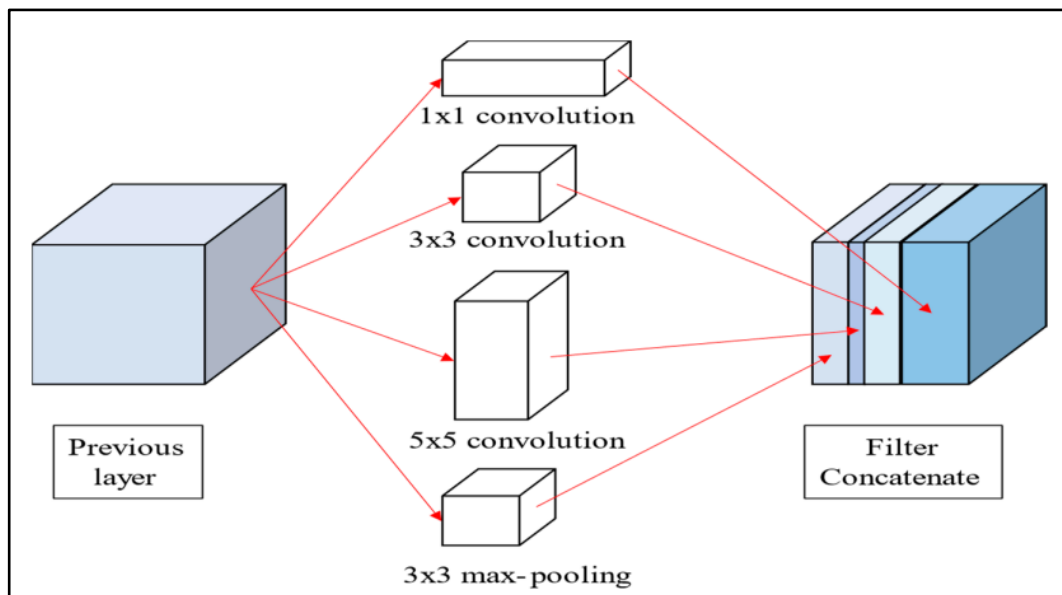


Figure 6: Typical inception module architecture

(Source: Mascarenhas and Agarwal, 2021)

The confusion matrix will be presented to analyse the general trends of the mistakes, namely, the overpredicted or underpredicting mistakes that may be further examined in detail. Through the consideration of the predicted outputs together with these performance measures, the strengths and weaknesses of the models will be established. In case of a mismatch in the above tasks, all variations or errors in classification will be followed up to get to the root cause of the problem, which could be problems with data quality or with the model (Mascarenhas and Agarwal, 2021). These results will help make further modifications, including model tuning or data enhancement, for higher classification effectiveness. These interpretations will give us ideas on the future development of deep learning models for medical image analysis.

Conclusion and Recommendation

The classification of breast histopathology images, using CNN and VGG16 models, would serve as a novelty as such ranges within the domain of utilizing deep learning models to automate classification in Medical Image Analysis. The study will use the “Breast Histopathology Image” dataset from Kaggle; The study shall use very strict approaches including EDA Data Remediation, and Model Development. Based on evaluation measures of the models, the findings of the study will offer an understanding of the effectiveness and capabilities of forms of deep learning methodologies in identifying cancerous cells in histopathology images. The results are expected to be useful in the existing literature on medical image classification especially in the diagnosis of breast cancer.

This is also suggested that the following points should be given a second thought to improve the performances of the model, including class imbalance problems of data augmentation strategies and improving the robustness of the model (Jiang et al. 2021).

Additional hyperparameters optimization, researching deeper structures of the deep learning network, and finding the opportunity to take advantage of transfer learning from existing models now also enhance the outcomes. Future work can also be undertaken using more diverse or larger data sets to improve the model's generalisability (Sriram et al. 2021). The utilisation of the explainable AI methods could add more detail to the way models give diagnosis adding to the efficacy of the automated systems in the medical environment.

Project Timeline

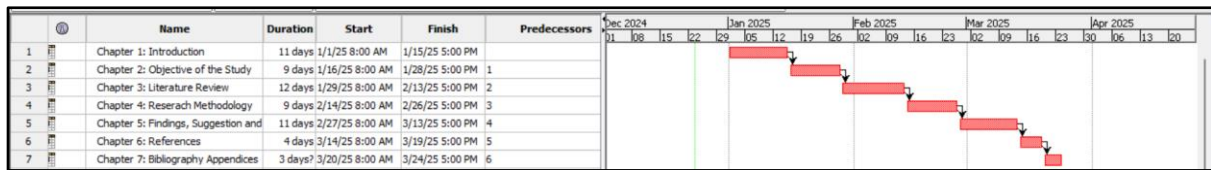


Figure 7: Timeline

(Source: Self-created)

References

- Burçak, K.C., Baylan, Ö.K. and Uğuz, H., 2021. A new deep convolutional neural network model for classifying breast cancer histopathological images and the hyperparameter optimisation of the proposed model. *The Journal of Supercomputing*, 77(1), pp.973-989.
- Desai, P., Pujari, J., Sujatha, C., Kamble, A. and Kambli, A., 2021. Hybrid approach for content-based image retrieval using VGG16 layered architecture and SVM: an application of deep learning. *SN Computer Science*, 2(3), p.170.
- Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J. and Maria Vanegas, A., 2020. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, 20(16), p.4373.
- Jiang, Z.P., Liu, Y.Y., Shao, Z.E. and Huang, K.W., 2021. An improved VGG16 model for pneumonia image classification. *Applied Sciences*, 11(23), p.11185.
- Mascarenhas, S. and Agarwal, M., 2021, November. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)* (Vol. 1, pp. 96-99). IEEE.
- Naveen, P. and Diwan, B., 2021, March. Pre-trained VGG-16 with CNN Architecture to classify X-Rays images into Normal or Pneumonia. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 102-105). IEEE.

- Shahidi, F., Daud, S.M., Abas, H., Ahmad, N.A. and Maarop, N., 2020. Breast cancer classification using deep learning approaches and histopathology image: A comparison study. *Ieee Access*, 8, pp.187531-187552.
- Sriram, G., Babu, T.R., Praveena, R. and Anand, J.V., 2022. Classification of Leukemia and Leukemoid Using VGG-16 Convolutional Neural Network Architecture. *Molecular & Cellular Biomechanics*, 19(2).
- Wetstein, S.C., de Jong, V.M., Stathonikos, N., Opdam, M., Dackus, G.M., Pluim, J.P., van Diest, P.J. and Veta, M., 2022. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Scientific reports*, 12(1), p.15102.