Graph-Based Convolutional Networks for Prostate Cancer Classification: Implementing Heterogeneous Information Networks and Deep Learning-Based Feature Extraction

Yash Bhargava
Student ID: x22220861
MSc Data Analytics
National College of Ireland

Abstract

Prostate Cancer is one of the most common types of cancer in men worldwide It is essential to detect and diagnose it before it may become severe. In recent times, technological advancement has helped us detect it in an early stage using machine learning, deep learning techniques and Artificial Intelligence. This research study aims to improve the detection and prostate cancer diagnoses by utilizing deep learning-based techniques and Heterogeneous Information networks (HINs) for the extraction of significant features and implement them with graph-based models such as Graph Attention Network (GAT), GraphSAGE, Graph U-Net and DCNN. Based on that we will evaluate the performance metrics such as Accuracy, Precision, F1-Score and AUC score and determine the effective feature extraction method with graph-based models to resolve the existing gaps in previous research studies.

Keywords— **P**rostate Cancer Detection, Heterogeneous Information Networks (HINs), Deep Learning, Graph Convolutional Networks, Classification

1 Introduction

Prostate Cancer is one of the most severe and deadliest diseases and the most common type of cancer among men in the US. It is a disease which can occur in the male prostate which is a walnut-shaped gland which is used to produce seminal fluid which can nourish and help sperm transportation. It is growing day by day which can cause men a serious problem and health issues. In Ireland, it is the second most common type of cancer and more than 2500 new cases of prostate cancer discovered. Many risk factors can cause prostate cancer such as the age of a male, family history, and diet also play an important role in lowering the risk of prostate cancer and weight of a male. This type of cancer is mostly detected in males aged 60 years or older. Early detection and diagnosis of Prostate Cancer is very essential as it may cause serious health issues such as urination problems in men. In men, the prostate is responsible for the production of prostate-specific antigen (PSA). The levels of PSA can be traced in our blood. By which high PSA levels show the condition of the prostate Janney et al. (2017). The detection of this type of cancer is crucial and the diagnosis of prostate cancer is successfully at an early stage. Researchers have proposed models for the classification and detection of prostate cancer using deep learning models such as Improved ResNet50, Convolutional Neural Networks (CNN), DenseNet, InceptionV3, VGG16, VGG19 and many more and they have also proposed their models by changing the baseline architecture of these above-mentioned models. Janney et al. (2017) has obtained the Region of Interest (ROI) using artificial neural networks (ANN) and then they have done segmentation. They mainly have proposed a model in which they have used the Radial Basis Function network with the Gaussian kernel for pattern region in the region of interest. In the above research, they have not explored graph representation-based learning and most of the researchers have done classification using deep learning models. They have done some hyperparameter tuning like changing the number of layers and model parameters. Still, I find some limitations in that these research studies don't explore graph-based representation learning and have not used any feature extraction techniques that have not been explored yet. Hence In this research, I will try to narrow this gap by implementing graph-based convolutional models and will implement these models and use 2 different feature extraction techniques to compare the performance metrics of the graph-based models.

1.1 Research Question

To move forward with the research, we should have a challenging research question to answer in this research study.
"How does the integration of features from heterogeneous information networks (HINs) impact the performance of graph-based convolutional network models as compared to transfer learning-based feature extraction techniques in the classification of prostate cancer from medical imaging data?"

1.2 Research Objective

In this research study, I will investigate and conduct a comparative analysis of graph-based neural networks by integrating them with features that will be extracted using Heterogeneous Information Networks (HINs) and deep learning-based models for the classification of Prostate Cancer MRI images and evaluate their performance based on Accuracy, Precision, F1 Score and ROC Score. This research will directly assist the healthcare sector such as medical professionals, patients, healthcare institutions researchers and developers to make informed decisions about the early detection and diagnosis of prostate cancer. With this many healthcare institutions can allocate their resources for the improvement of their healthcare delivery. It can also help technology specialists to develop various AI tools for innovation in the healthcare sector.

This research proposal has been broken down into many sections. In the next section 2, I have defined the summary of previous research works on this domain subject. In Section 3, I have defined my research methodology and proposed implementation and specification. In Section 4, I have defined the ethical consideration for this research to be carried out successfully, Project Management plan. Finally, I have defined the conclusion of this research work.

2 Literature Review

In this section of the study, we will review the multiple research papers that help in the inclusion of the healthcare domain knowledge and study various methodologies used by researchers. This section of the proposal is divided into the following subsections:

2.1 Prostate Cancer using Machine Learning Techniques

The research was done by Dubey & Kumar (2023) the authors applied principal component analysis (PCA) to reduce the dimensions and as a data preprocessing and then classified the transformed dataset with k-Nearest Neighbours classifiers which have an accuracy of 97.35% with the number of neighbours taken as 50 and 96.42% with the number of neighbours as 100. Another research was conducted by Duenweg et al. (2023) in which the study compares the 2 machine learning methods for the classification of cancerous regions on digitized histology from 47 Prostate Cancer patients. In this study the patients were divided into 2 sets consisting of the training set of 31 patients and the testing set of 16 patients used in which training set was labelled as Cohort A of n= 9345 tiles and the testing set was labelled as Cohort B of n= 4375 tiles and the training set was trained using ResNet model the glands of these tiles were classified to distinguish between the tumours. The ensemble model and ResNet model have an accuracy of 89% and 88% respectively. In this research, the dataset size used was very small only 47 patients and the use of graphical convolutional networks can give better accuracy and results. Another research was conducted by Iqbal et al. (2021) the authors used LSTM and ResNet101 without the hand-crafted features and it was tuned. The results of these models were compared with hand-crafted features integrated with traditional machine learning classifiers i.e. Support Vector Machine (SVM) with Gaussian Kernel, k-nearest neighbour with cosine similarity, Naïve Bayes, Decision Tree (DT) and RUSBoost tree. For the validation of these models, 10 cross-fold validation was also used. The result of this study shows that the ResNet101 model outperformed all the other models with an accuracy of 100% and an AUC value of 1 which shows that ResNet101 was the best predictor for the classification of prostate cancer. In this study, ResNet101 was used without the extraction of hand-crafted features and hence it is giving the best performance among them it was not accurate and hence we can extract significant hand-crafted features and the performance metrics can be more accurate and we can also employ the utilization of graph-convolutional networks (GCN) and we can use Heterogeneous Information Networks to extract significant features.

2.2 Prostate Cancer Classification Deep Learning and Neural Networks

Another research was done by Janney et al. (2017) in which the paper was about the detection of prostate cancer with artificial neural networks [ANN]. This method was employed in this research to detect early-stage prostate carcinoma. The MRI images were pre-processed to reduce the effects of noise from the images and the region of interest was found with ANN and classification was done based on it. The main aspect of this research study was that every region of the cancerous cell was as malignant or abnormal tissue by the utilization of the Radial Basis Function network with Gaussian kernel for recognition of the pattern. Another research was done by Talaat et al. (2024) in which the utilization of deep learning algorithms for the classification of prostate cancer detection in which the author had employed the modified ResNet50 integrated with R-CNN and optimizers and the comparative analysis was done with normal ResNet50 and VGG19 and it performed better as compared to the proposed model and achieved the accuracy of 95.24%. However there might be a slight challenge that they don't explore other deep learning models and the proposed model was judged based on accuracy and other performance metrics, but this proposed model can work better if it can be integrated with graphical convolutional networks where the nodes of the graphs can be represents patients' information and graph-based techniques can be used to determine the patterns and large datasets.

Another research was done by Yuan et al. (2019) in which the authors did prostate cancer segmentation of Magnetic Resonance (MR) images for which they employed an Encoder-Decoder Densely Connected Convolutional Network (ED- DenseNet). In this model, the authors have used two interconnected pathways, a dense encoder pathway and a dense decoder pathway which were used to predict the final segmentation at the pixel level. The reason behind using the Densely Connected Convolutional Network was to preserve the maximum information in the Magnetic Resonance (MR) images by using the densely connected structure. However, the authors also used a loss function for the encoder-decoder error and prediction error for the optimization of learning features and classification results. The results of this study claim that ED-DenseNet can get the highest accuracy when compared to other deep learning models. In this research study, the authors have done very little comparative analysis so we cannot interpret the results. Another research was done by (Gavade et al. 2023) in which the authors employed multi-parametric magnetic resonance imaging integrated with deep learning models for the segmentation and classification of various types of cancer. In this research study, the authors have used 4 pipelines i.e. Semantic DeepSegNet with ResNet50, DeepSegNet with recurrent neural network (RNN), U-Net with RNN and U-Net with long-short-term memory (LSTM). Every segmentation model was paired with various classifiers for the evaluation of performance in which the authors found out that the pair of U-Net and the LSTM model outperformed all the combinations. In this research, there were 4 different pipelines were used which increased the complexity of this research study. We can explore heterogeneous information networks based on graph convolutional networks that can increase interpretability and allow healthcare clinicians to better understand the results of cancer classification and HIN-based models can be clinically essential for the findings and more effective and efficient.

Another research was done by Hong et al. (2023) the authors tried to find the feasibility of the deep learning models depending upon the apparent diffusion coefficient (ADC) maps for the differentiation between the cancerous and non-cancerous cells in which the cancerous cells have Gleason score ¿= 7. In this research study, the data used has been collected from 149 patients who have diagnosed with Prostate Cancer were collected. Out of which the labelled data has 148 images of GS6 and 580 images for GS7 were used for tumour segmentation. Overall, for the classification of the images, U-Net and DenseNet were used. The internal and external validation accuracies were 73% and 75%. In the end, the authors concluded that the differentiation of CSC from non-CSC was feasible using deep learning algorithms.

2.3 Prostate Cancer Classification using Artificial Intelligence and Convolutional Neural Networks

Another research was done by Sethi et al. (2023) the authors developed a new machine learning and deep learning approach i.e. Gene Expression Data Analysis using Artificial Intelligence for Prostate Cancer Diagnoses (GEDAAI-PCD). In this gene expression data was normalised into a uniform format and then the Long Short-Term Memory–Deep Belief Network (LSTM-DBN) model was used for the classification of Prostate Cancer. For hyperparameter tuning wild horse optimization was used so that the performance of the model was improved. In this research, the authors integrated the Long-Short Term Memory Deep Belief Network (LSTM-DBN), the specificity of the GEDAAI-PCD system was 98.75% which was higher than the other applied models which shows the best performance among them. However, the model was trained on a small dataset of only 102 instances. Finally, the research was conducted by Mosleh et al. (2022) in which the authors implemented an automated process for the detection of Prostate Cancer by the utilization of MRI images with the CNN. The authors have implemented the pre-trained transfer learning models i.e. Inception-v3, Inception-v4, Inception-Resent-v2, Xception, and PolyNet. The use of transfer learning models for the classification of the MRI images into two sets: positive and negative results. The results of this research work for all the models were extremely good and PolyNet outperformed all the other models with an accuracy of 99.34%. But the dataset used in this research was small i.e. 1624 MRI images only. However, we can integrate graphical convolution networks on a large dataset so that the results are easily interpreted and some of the performance metrics are improved.

2.4 Conclusion of the Literature Review

So, after reviewing all the papers I have seen that there can be many different ways and methods for the classification of prostate cancer such as traditional machine learning models, deep learning techniques and neural networks as well but they have not explored the graph based convolutional models and not utilised any feature extraction techniques which can impact on the performance of the models and along with graph-based convolutional model so that this aspect of research is remaining and I think that it is very important to explore that graph based models for the classification of Prostate Cancer as it will provide assistance to medical professionals and many medical organizations te make quick decision for the early detection and diagnosis of the dadle disease among men.

3 Research Methods & Specifications

In this research, I will be using the Knowledge Discovery from Databases (KDD) methodology as it is a very robust and easily interpretable methodology which can assist in the detection and classification of Prostate Cancer. The data mining technique used in this research is the classification of Prostate Cancer as the labelled class as "isup_grade" and for the classification, I will use Heterogeneous Information Networks (HINs) and deep learning techniques for the extraction of features from MRI images and graph-based convolutional network (GCN) models will be utilized to detect and classify Prostate Cancer.

3.1 Proposed Methodology

- **Data Selection-** The first step of the KDD methodology is the selection of data I have chosen my Prostate Cancer MRI images dataset for this research study from Kaggle which is a huge repository of datasets worldwide. The dataset has 10516 images of 512*512 of Prostate Cancer in TIFF format and labels of each image in CSV format. There are also label masks of the images in the TIFF format. The CSV file has 10616 rows and 4 columns labelled for each image.

- **Data Preprocessing-** This is the second step of the KDD methodology in which I will be removing noise, data augmentation, resizing and resampling the images and normalising the MRI images. Also, I will do feature extraction from the images using Heterogeneous Information Networks and deep learning techniques. In this step, the data will be cleaned as per the requirement to train the model for classification and detection of Prostate Cancer.

- **Data Transformation-** This is the next step of KDD methodology in which I will transform my Prostate Cancer Images dataset by applying some geometric transformations such as scaling and flipping the image and this will further assist my graph-based models so that my model doesn't

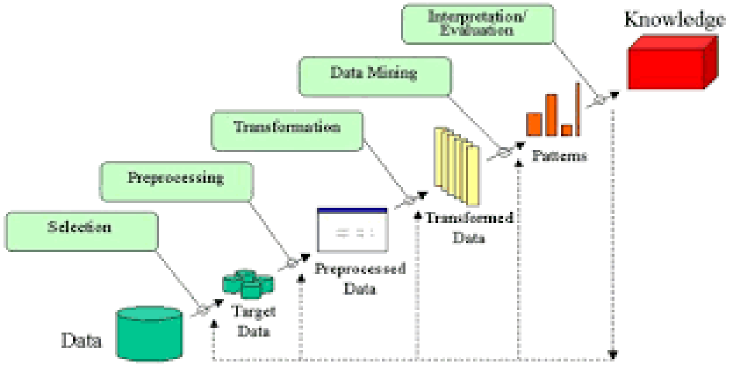

Figure 1: KDD Methodology

overfit and performs better and efficiently. After applying the transformation to the dataset, it will become arbitrary and not like the original dataset.

- **Data Mining-** In this step of this methodology, I will divide my datasets into two sets 70% as training set images and 30% of the dataset into a test set. As data mining is the main step of this methodology, I will train the model based on the training set based on the features extracted from the images and classify prostate cancer images by making use of graph-based models such as Graph Attention Network (GAT), GraphSAGE, Graph U-Net and DCNN.

- **Results Interpretation-** In this final step, after the training and testing of the model, I will evaluate the key performance metrics i.e. Accuracy, Precision, Recall, and ROC Score and based on that I will compare the results of each model based on the features extracted from Heterogeneous Information Networks and using deep learning techniques.

3.2 Design Specification

In this research study, firstly I downloaded the PANDA challenge Prostate Cancer dataset, which was publicly available, then I loaded and preprocess the images and constructed a Heterogeneous Information Network (HIN) which is used to extract the features from the 11000 images and these features will be integrated with the network. In my next step, I will implement graph-based convolutional network models such as Graph Attention Network (GAT), Graph Sample and Aggregation (GraphSAGE), Graph Isomorphism Network (GIN), Graph U-Net and Diffusion Convolutional Neural Network (DCNN) and I will do a comparison of the performance metrics of these models and try to get better performance from the models and this research will fit the gap of the previous research as these graph-based models have not been implemented before as per my review of a research paper so I will extract feature using deep learning based feature extraction technique using transfer learning and Heterogeneous Information Networks (HINs) and then do classification using above models on both the features that are extracted using those feature extraction techniques and finally present the results by evaluating the performance metrics of the models which shows that which feature extraction techniques works better with graph-based network models.

3.3 Proposed Implementation

To implement this research work, the focus is also to utilize the performance of the graphical models from the earlier work, for this purpose the steps need to be taken when we apply the above-mentioned methodology and the system configuration and programming tools for the construction of Heterogeneous Information Networks, implementing deep learning techniques and graph-based models.

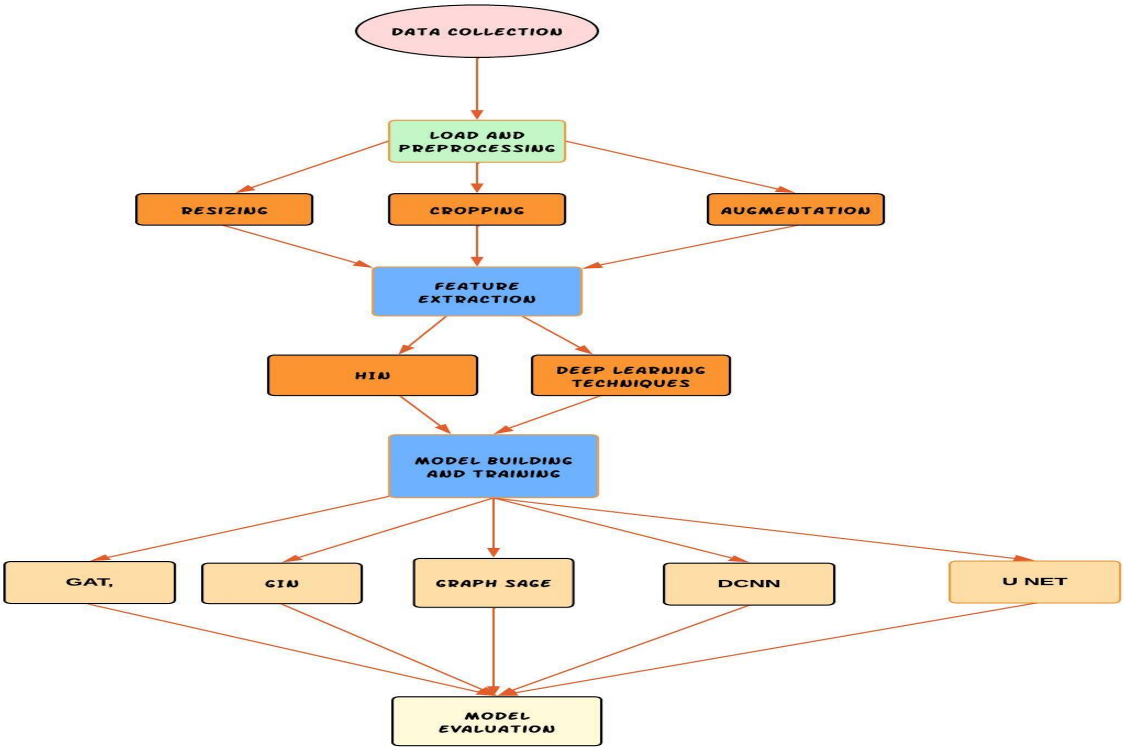

Figure 2: Design Specification

3.3.1 System Configuration

The proposed implantation of the models will be executed in the following System Configuration:

1. **Device:** HP Laptop 15s-du3047TX

2. **Operating System:** Windows 11 x64-bit

3. **RAM:** 8.00 GB

4. **CPU:** Intel(R) Core (TM) i5-1135G7 @ 2.40GHz

5. **GPU:** Intel(R) Iris(R) Xe Graphics, NVIDIA GeForce MX350

3.3.2 Tools and Libraries

In this research work, I will use Jupyter Notebook, Google Collab, Python programming Language, sckit learn library, PyTorch, TensorFlow, NetworkX, NumPy, Pandas, OpenCV and DeepGraph (DGL) library.

4 Ethical Considerations of the Research

Ethics are necessary for the creation of projects and planning. Considerations include social attitudes, resources, private databases, human participation, and environmental impact. The owner must permit using private assets or private secondary data for preparation, organizing, and implementation. The dataset from this study is easily accessible for educational and research purposes. Users are not required to register or pay to use the dataset.

5 Project Plan

Planning is a very important part before carrying out any research project so that the execution of each task. For this research work, there is a 12-week schedule which will begin from 16th May 2024. For the first 12 days, I will conduct background research on this subject till 28th May 2024. Next, I will conduct a literature review for 14 days for about 15-20 research papers till 12th June 2024. I will further collect the dataset and preprocess the dataset for 20 days till 04th July 2024. Then I will perform some kind of Data Transformation and feature extraction for 7 days till the 12th of July 2024. Then I will further build, train and evaluate key performance metrics which will take 15 days till 29th July 2024. In between the above-mentioned dates, I will also attend the meetings with my mentor and finally, I will write my report and do the documentation for the research work at the end I will submit the final report draft. For the reference, the Gantt chart is shown in Figure 3.

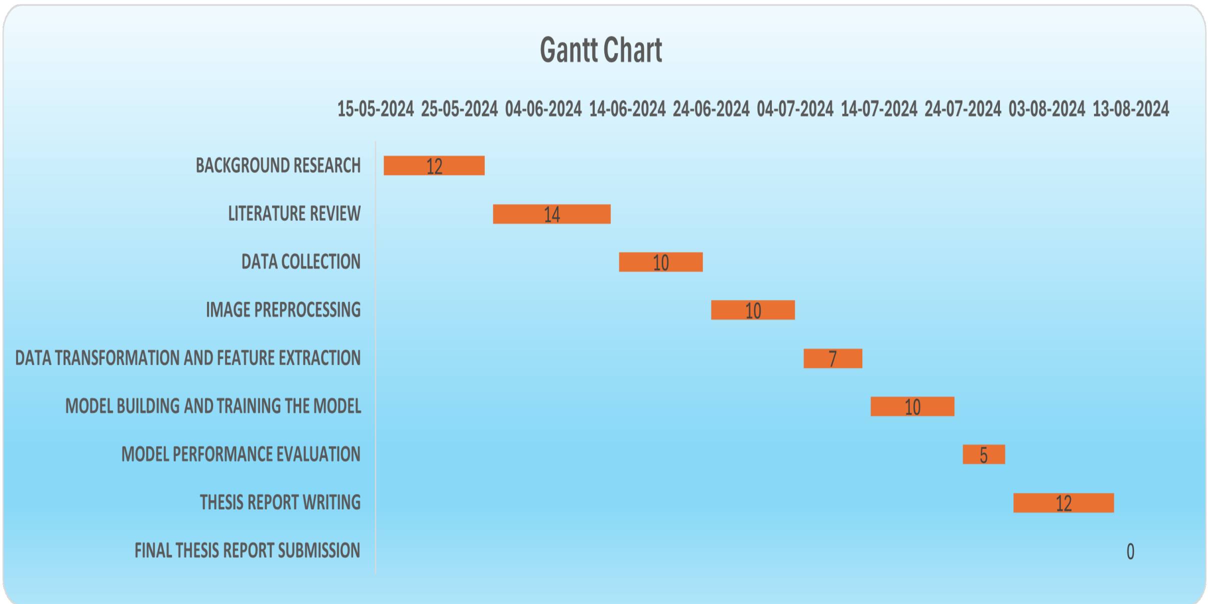

Figure 3: Project Plan for Research Work

6 Conclusion

Finally, the main aim of this research project is to investigate the best combination of features that were extracted from the construction of Heterogeneous Information Networks (HINs) and transfer learning-based techniques with the graph-convolutional models i.e. Graph Attention Network (GAT), Graph Sample and Aggregation (GraphSAGE), Graph Isomorphism Network (GIN), Graph U-Net and Diffusion Convolutional Neural Network (DCNN) and calculate the key performance metrics such as Accuracy, Precision, Recall, F1 Score and ROC AUC Score. This research study proposal provides a comprehensive and unique approach to prostate cancer diagnosis by combining modern techniques in image processing, deep learning, and graph-based modelling. The results of this research have been expected to have an important effect on the healthcare sector because they provide information about the most successful ways to use MRI data in prostate cancer detection and treatment strategies. This research study also benefits many clinicians, technology experts, healthcare professionals and organizations for the detection of Prostate Cancer and diagnosis in its early stages so that many patients don't die because of cancer.

References

Dubey, P. & Kumar, S. (2023), Pca-knn classifier for improved diagnostic accuracy: A study on prostate cancer, *in* '2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)', pp. 1–5.

Duenweg, S. R., Brehler, M., Bobholz, S. A., Lowman, A. K., Winiarz, A., Kyereme, F., Nencka, A., Iczkowski, K. A. & LaViolette, P. S. (2023), 'Comparison of a machine and deep learning model for automated tumor annotation on digitized whole slide prostate cancer histology', *Plos one* **18**(3), e0278084.

Gavade, A. B., Kanwal, N., Gavade, P. A. & Nerli, R. (2023), 'Enhancing prostate cancer diagnosis with deep learning: A study using mpmri segmentation and classification', *arXiv preprint arXiv:2310.05371* .

Hong, S., Kim, S. H., Yoo, B. & Kim, J. Y. (2023), 'Deep learning algorithm for tumor segmentation and discrimination of clinically significant cancer in patients with prostate cancer', *Current Oncology* **30**(8), 7275–7285.

Iqbal, S., Siddiqui, G. F., Rehman, A., Hussain, L., Saba, T., Tariq, U. & Abbasi, A. A. (2021), 'Prostate cancer detection using deep learning and traditional techniques', *IEEE Access* **9**, 27085–27100.

Janney, J. B., Christilda, J. J., Mary, S. S. & Haritha, D. (2017), Early diagnosis of prostate cancer using image processing techniques, *in* '2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)', pp. 2868–2871.

Mosleh, M. A., Hamoud, M. H. & Alsabri, A. A. (2022), Detection of prostate cancer using mri images classification with deep learning techniques, *in* '2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)', IEEE, pp. 1–6.

Sethi, B. K., Singh, D., Rout, S. K. & Panda, S. K. (2023), 'Long short-term memory-deep belief network based gene expression data analysis for prostate cancer detection and classification', *IEEE Access* .

Talaat, F. M., El-Sappagh, S., Alnowaiser, K. & Hassan, E. (2024), 'Improved prostate cancer diagnosis using a modified resnet50-based deep learning architecture', *BMC Medical Informatics and Decision Making* **24**(1), 23.

Yuan, Y., Qin, W., Guo, X., Buyyounouski, M., Hancock, S., Han, B. & Xing, L. (2019), Prostate segmentation with encoder-decoder densely connected convolutional network (ed-densenet), *in* '2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)', IEEE, pp. 434–437.