International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I12P118

Volume 73 Issue 12, 218-228, December 2025

Original Article

A Comparative Study of Prostate Cancer Classification Using MRI Images with a Machine Learning Approach for Early Diagnosis

Neny Rosmawarni[1*], Zeratul Izza Mohd Yusoh[2], Yun Houy Choo[3], Thoyyibah[4], Karunia Agustiani[5], Nadra[6]

[1,5]*Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia.*
[1,2,3]*Informatics, Universiti Teknikal Malaysia Melaka, Malaysia.*
[4]*Information System Management Department, BINUS Graduate Program- Master of Information System Management University of Bina Nusantara, Jakarta, Indonesia.*
[6]*Radiology Department, Gatot Subroto Army Hospital (RSPAD), Jakarta, Indonesia.*

[1]*Corresponding Author : nenyrosmawarni@upnvj.ac.id*

Received: 23 May 2025 Revised: 08 November 2025 Accepted: 25 November 2025 Published: 19 December 2025

Abstract - *Prostate cancer is one of the most common malignant cancers worldwide. Early detection and diagnosis are essential for treating this Cancer. This study uses features extracted from the Grey Level Co-occurrence Matrix (GLCM) with the Extreme Gradient Boosting (XGBoost) classifier to improve prostate cancer classification using MRI images, with training, validation, and testing. 961 public MRI images consisting of 424 cancerous and 537 non-cancerous images were used. GLCM was used at four angles (0°, 45°, 90°, and 135°) for several texture features: correlation, energy, and homogeneity. The experimental results on the training data achieved an accuracy of 99.8%; on the validation data, the accuracy reached 63.2%; and on the testing data, the precision was 71.6% and the recall was 70.37%. The accuracy results show that the GLCM with the XGBoost model is very effective at capturing discriminative features and achieving balanced classification performance. The proposed model presents a promising foundation for developing automated, data-driven tools in early prostate cancer detection. Future research will focus on hyperparameter tuning, data augmentation, and regularisation to further improve model generalisation and clinical applicability.*

Keywords *- Prostate cancer, MRI Images, GLCM, XGBoost, Machine Learning, Classification.*

1. Introduction

The prostate is the largest accessory gland in men and is part of the male reproductive system. The prostate gland is located below the bladder and surrounds the urethra, which functions to complete the secretion of semen and keep sperm alive [1]. The prostate gland can be affected by Cancer. Prostate cancer is a condition in which prostate cells continue to grow and cause pain, as well as obstruction when urinating for sufferers. One Cancer is common in most older men and has the second-highest number of sufferers in the world, according to a report from the Cancer Agency, namely IARC, where the number of cancers with the fifth-highest score in Indonesia can be seen in Figure 1. In 2022, based on this data source, the WHO recorded 13,130 cases of Cancer with a death rate of 4,860 cases. With the increasing number of older adults in Indonesia, there is a growing risk factor for prostate cancer, which is often diagnosed at an advanced stage due to late detection [2]. Therefore, early detection is necessary so that patients can receive appropriate medical treatment immediately, increasing the chances of recovery. One of the primary methods to detect prostate cancer is through a medical examination in the form of an MRI scan, which a radiology specialist performs. MRI is commonly used for diagnosis, therapy planning, and evaluation of treatment response [3].

In addition, several previous studies have not compared the results of algorithm accuracy on this data. This is interesting because it not only uses a new approach but also compares the advantages and limitations of each model. So the results can be further utilised in technology-based analysis, including Machine Learning.

Classification is used to group datasets into specific categories based on patterns learned from training data. In MRI image data, Classification can be performed directly or first using extracted data to recognise the various objects being analysed. Feature extraction aims to identify key features in an image, while classification groups objects based on these features. Thus, combining these two can help detect prostate cancer more accurately and efficiently.

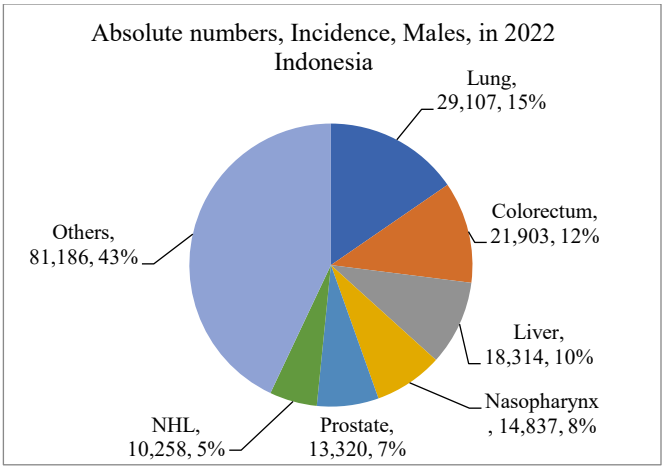


Fig. 1 Cancer data from the International Agency for Research on Cancer (IARC)

Various studies on classified images using multiple algorithms. Research on MRI image classification was conducted by [4], who used 7023 brain MRI images from 4 tumour classes and the Random Forest method to classify tumour type, achieving an accuracy of 91%. Another study related to prostate cancer [5] used tabular data with 100 records and applied the Naïve Bayes classification algorithm, which achieved 80% accuracy, while KNN achieved 90%. Next is the research by [6], which classifies heart disease using 4238 records and 16 attributes via the boosting algorithm, namely XGBoost.

The accuracy was 84.6% for XGBoost and 91.6% for XGBoost by applying Adaptive Synthetic Sampling. Based on previous research, this study will develop research related to MRI images to classify prostate cancer based on the extraction carried out on the scan features with a grey level called GLCM and the use of boosting-based Classification, namely Xgboost. The research will be conducted using feature extraction from MRI images and the XGBoost classification method. So this study is expected to provide good accuracy results in the Classification of prostate cancer and can help patients to immediately get appropriate medical treatment so that prostate cancer can be recognised at an early stage. Patients will have a higher recovery rate. The novelty of this study lies in integrating GLCM-based feature extraction with the XGBoost classifier for prostate MRI image classification, a task that has not been extensively analysed in previous work. This approach is expected to improve accuracy and computational efficiency in early prostate cancer detection.

2. Related Work

2.1. Previous Studies

In support of this research, the researchers conducted a literature review on a specific topic. This can increase researchers' insight into the features extracted from the grey level, namely the GLCM, and the Extreme Gradient Boosting (XGBoost) classifier. Several previous studies related to this topic have been identified as sources and supporters in the implementation of the research compiled in Table 1.

Table 1. Literature review

No	Title	Method	Results
1	Brain Tumour Classification Based on MRI Images Using Random Forest Classifier with GLCM Feature Extraction [4]	GLCM Feature Extraction and Random Forest Classification Algorithm	Utilised MRI image data from Kaggle, consisting of 7,023 MRI images across four classes. This study applied GLCM feature extraction. Classification using Random Forest with n_estimators=140 achieved an accuracy of 91%.
2	Prostate cancer analysed using KNN and Naïve Bayes [5]	KNN and Naïve Bayes classification	Used data from Kaggle containing 100 records with nine attributes. 80% for Naïve Bayes, while K-Nearest Neighbour with K=1-7 achieved the highest accuracy of 90% at K=5.
3	Cardiovascular Disease Classification Using Adaptive Synthetic Sampling and Extreme Gradient Boosting Algorithm [6]	Extreme Gradient Boosting Classification Algorithm and Adaptive Synthetic Sampling Method	Used Kaggle data with 4,238 records and 16 attributes. After applying ADASYN, the dataset increased from 4,238 to 7,227 records to address class imbalance. Classification using Extreme Gradient Boosting achieved an accuracy of 84.6%, which improved to 91.6% after ADASYN was applied.
4	Analysis on sentiment and usage boosting through XGBoost [7]	Extreme Gradient Boosting Classification Algorithm	Utilised web-crawled data consisting of 2,243 tweets and 305 comments collected between March and July 2022. Classification using Extreme Gradient Boosting achieved an accuracy of 91.09% with an 80:20 data split.
5	MRI Brain Tumour Classification Using Grey Level [3]	GLCM Feature Extraction - Naïve Bayes, C4.5, Neural Network Algorithms	Utilised MRI image data from Kaggle with 2,870 MRI images. This study applied GLCM feature extraction and compared multiple classification methods. Naïve Bayes achieved an accuracy of 96.8%, C4.5 achieved 41.5%, and NN achieved 38,25%.

Some previous studies have identified gaps in the application of GLCM and XGBoost-extracted features for prostate cancer image classification, as these areas have not been analysed in depth in prior experiments. This study aims to bridge this gap by integrating GLCM feature extraction with XGBoost classification to achieve high accuracy in classifying MRI prostate cancer images.

2.2. Theory

2.2.1.Prostate Cancer

The prostate is a small gland in the male reproductive system. It is located just below the bladder and surrounds the urethra. The prostate's role is to produce fluid that becomes part of semen and helps in maintaining the viability of sperm [1]. Prostate cancer is the disease with the second-highest cancer rate in men. Prostate cancer is a condition where cells in the prostate gland continue to experience unnatural growth. This unnatural growth and development can cause obstacles and pain when urinating in patients [8]. The risk factors behind the occurrence of prostate cancer include age, family history, obesity, diet with low fibre foods and high animal fat, smoking, drinking alcoholic beverages, and disease, especially in the prostate. Prostate cancer often occurs in older men; it is rarely found in men aged 50 years and under. However, due to an unhealthy lifestyle, there are also prostate cancer patients under the age of 50 [9].

2.2.2. Magnetic Resonance Imaging (MRI)

An MRI scan is a medical imaging technique that produces detailed images of internal organs and the human body using radio waves and magnetic fields. This scan does not use X-rays or radioactive materials [10].

2.2.3. Image Grayscale

Grayscale images are images resulting from the conversion of colour images to grayscale. A grayscale image is described as a two-dimensional array. The image's grey-level intensity is represented as a two-dimensional array, with corresponding positions. Each pixel in a grey image has a colour spectrum ranging from white to black. Generally, this colour range is used in the medical world. For example, in the CT-Scan or MRI process [11].

The grayscale value can be obtained by using (1) to convert the RGB image value to the grayscale image model [12].

$$I_{BW}(x,y) = (I_R \times 0.2989) + (I_G \times 0.5870) + (I_B \times 0.1140) \quad (1)$$

2.2.4. Image Segmentation

The technique of dividing an image into segments or regions is called image segmentation. This segmentation is a process of analysing and describing image objects by separating them from their background. Various applications can be used to analyse segmented objects [13]. Thresholding is a method of converting grayscale images into binary images. In general, pixel images with objects have a value of 1 and images with backgrounds have a value of 0 [14]. Automatically analyse images with backgrounds and objects with black and white values [15]. A segmentation technique that uses a specific area is called ROI (Region of Interest). This ROI is used to reduce the complexity of focusing on relevant images in image processing [16].

2.2.5. Features Extracted with Grey Level

Specific patterns are formed from the texture of digital images. Pixel interactions are used to visually examine patterns by identifying and analysing image characteristics using existing feature extraction [17]. The method most often used is the Grey Level Co-occurrence Matrix (GLCM). This method considers the spatial distance at a given angle, representing the relationship between neighbouring pixels in digital images, using the co-occurrence matrix [18]. Homogeneity, correlation, contrast, and energy are the GLCM features used.

Contrast is: a measure of the differences in pixel intensities between pixels and their surroundings across images, capturing variations in grayscale intensity using the formula (2).

$$Contrast = \sum_{i,j}(i-j)^2 P_{(i,j)} \quad (2)$$

Correlation: Correlation represents the degree of linear relationship between pairs of pixels in an image with formula (3).

$$Correlation = \frac{\sum_{i,j}(i-\mu_i)(j-\mu_j)P_{(i,j)}}{\sigma_i \sigma_j} \quad (3)$$

Energy: Energy measures uniformity and is related to the number of variations in grayscale intensity, as shown in formula (4).

$$Energy = \sum_{i,j} P_2^d\,(i,j) \quad (4)$$

Homogeneity: Homogeneity evaluates the similarity and proximity of intensity variations in an image with formula (5).

$$Homogeneity = \sum_{i,j} \frac{pd(i,j)}{1+|i-j|} \quad (5)$$

2.2.6. Data Normalisation

Data preprocessing is an essential stage before classifying. A critical step toward the end goal is data normalisation during preprocessing. Data refers to the transformation of data based on specific statistical properties. This process is essential because it completes the efficiency process at the model training stage, making it faster [19]. Min-Max normalisation is a data normalisation technique. This method applies the original data that has previously been

transformed and balances the normalisation process across the stages before and after normalisation, using predetermined lower and upper limits [20] as in formula (6).

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} (New_{max} - New_{min}) \quad (6)$$

2.2.7. Data Imbalanced

Data imbalance refers to a condition in which datasets are not evenly distributed across all classes. Typically, one or more classes have much smaller samples than the others. Classification performance can be affected by imbalance [21]. Random undersampling is a technique that randomly selects and removes samples from the majority classes in the training data. As a result, the number of large datasets in the class is reduced, thereby balancing the dataset and improving model performance [22].

2.2.8. Extreme Gradient Boosting (XGBoost) Algorithm

XGBoost is an algorithm for the machine learning part of the ensemble that uses a boosting-based decision tree. XGBoost improves on traditional gradient boosting by integrating decision trees with gradient boosting. This allows the algorithm to be highly scalable, efficient, and flexible in handling both classification and regression tasks. Additionally, XGBoost is recognised for its ability to handle large volumes of data and its speed when managing complex datasets [6]. In binary Classification, the objective function of XGBoost is designed to measure prediction errors and iteratively update the model to enhance accuracy. This function consists of two key components: the loss function, which evaluates prediction errors, and the regularisation term, which helps prevent overfitting [23].

2.2.9. Confusion Matrix

A confusion matrix is a two-dimensional table that compares actual and predicted values, describing a model's performance on test data. According to Kantardzic, this is useful for evaluating classification model performance and measuring how well the model classifies data [24] in classification evaluations. The four parts of matrix confusion include TP, which is true positives, meaning that when the actual class is positive, the model correctly predicts the positive class. TN or Negative True: when the actual class is positive, the class prediction results are correctly negative. FP or False Positive, which occurs when the actual class predicted by the model is incorrectly classified. FN or False Negative occurs when a class is actually positive, with a model that incorrectly predicts it as negative. Minimising TP and TN will result in a well-performing Model with a record of high TP and TN values—Matrix Convergence and metric evaluation results in model performance metrics such as accuracy, precision, and recall.

Accuracy: The entire model is correctly predicted using the formula (7).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (7)$$

Precision: The prediction results that were assessed positively from the evaluation of many positive cases were used with the formula (8).

$$Precision = \frac{TP}{(TP + FP)} \quad (8)$$

Recall: The identification results are correct if the proportion of cases that test positive is determined accurately using the formula (9).

$$Recall = \frac{TP}{(TP + FN)} \quad (9)$$

3. Materials and Methods

The data collection and preparation process for the study's Classification of MRI images of the prostate requires precision and significant computational resources, especially during image processing and the extraction of key characteristics using appropriate methods. Therefore, obtaining accurate information depends heavily on effective data, feature extraction, and classification methods. The design of these models is quite complex, as it requires selecting optimal parameters to improve classification performance.

In this study, the Grey Level Co-Occurrence Matrix (GLCM) method was used to extract features and identify texture characteristics of prostate MRI images. After that, the Boosting classification model, namely XGBoost, was developed and optimised to improve the accuracy of detecting prostate cancer. As shown in the experimental results, the applied method produces a reliable model for distinguishing MRI images with and without indications of prostate cancer (Figure 2). Each research stage in the chart above plays a vital role in ensuring that the data used are relevant to the research objectives and that the methods used are appropriate for answering the research questions. The further explanation of each stage illustrated in the chart above is as follows:

3.1. Problem Identification

At the problem identification stage, the researchers identified problems that became the focus of research on MRI images of prostate cancer, Grey Level Co-Occurrence Matrix feature extraction, and the Extreme Gradient Boosting classification algorithm. Therefore, the researcher can describe the problems to be studied related to the Classification of MRI images of prostate cancer and determine the objectives and results to be achieved in the study.

3.2. Literature Review

After the problem identification stage and the determination of the objectives and results to be achieved in the research, the next stage is to conduct a literature review.

Literature review is a process of reading references in the form of books, journals, or other relevant research on the topic of discussion, namely the Classification of MRI images of prostate cancer using Machine Learning algorithms, the extraction of texture characteristics, and other information related to and supporting the research.

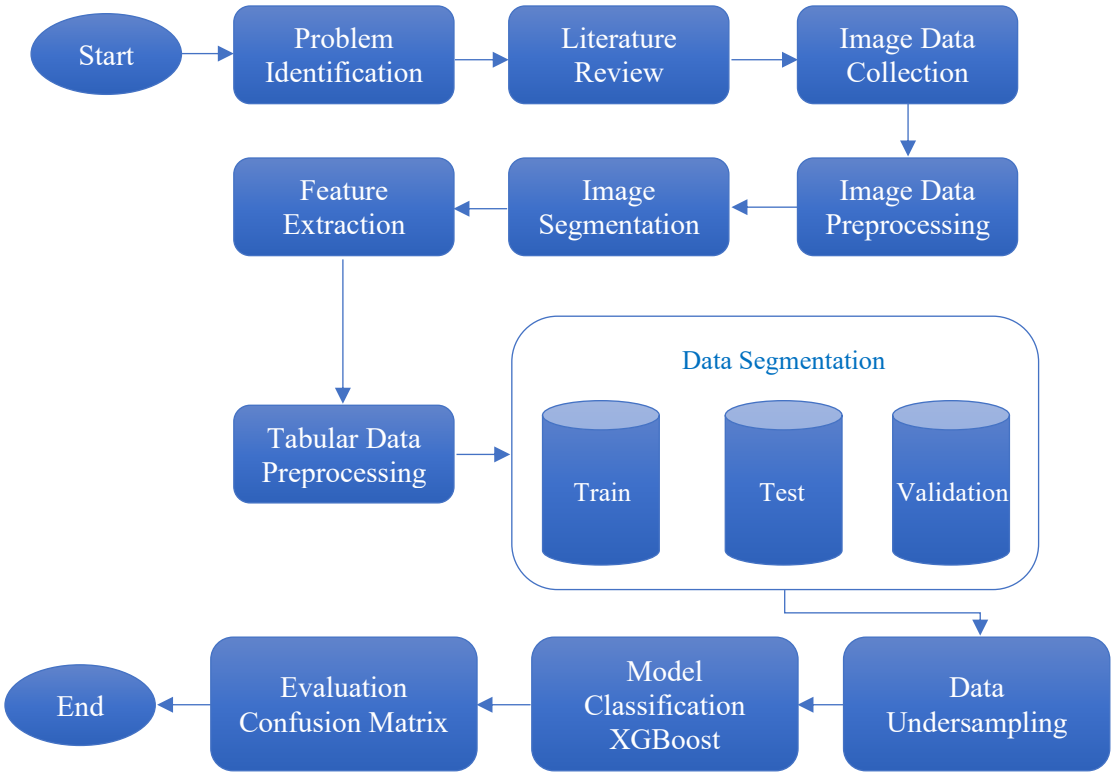


Fig. 2 Flowchart research

3.3. Data Collection

In the data collection stage, researchers collected data from open-source datasets on prostate MRI images, including Google Open Source and FastMRI [25]. The collected image data includes 537 non-prostate Cancer MRI images and 424 prostate cancer MRI images. The collected images are then converted and organised based on the presence of Cancer identified in each image. Data conversion involves converting the DICOM file format to the JPEG/JPG format.

3.4. Image Data Preprocessing

Image preprocessing is done to prepare the MRI image data before feature extraction. This image preprocessing stage includes the following digital image processing methods:

a) Image resizing is the process of reducing an image's size without cropping the original for consistency. So that the image remains in the initial State with the new size. In this research, the original image will be resized to 224x224 pixels.
b) Image grey scaling is the process of converting an image from the RGB colour space to a grey image. At this stage, the three colour space channels will be combined to form a single colour space with values ranging from 0 to 255, where the intensity of each channel is calculated using the corresponding red, green, and blue equations. The red channel is 0.2989, the green channel is 0.5870, and the blue channel is 0.1140. Besides that, this is also prepared for the image in the following process.
c) Image enhancement at the preprocessing stage is achieved by reducing noise with a median filter and sharpening contrast using the contrast-stretching method.

3.5. Image Segmentation

The segmentation method chosen in this research is Otsu thresholding, and the segmented mask is obtained using the ROI method. Where each stage has the following explanation: segmentation performed:

a) Thresholding with Otsu's Method is an automated image segmentation method that determines the optimal threshold based on the histogram distribution of pixels in a grayscale image by minimising the within-class variance of the two groups. So that all pixels in the image that have intensities below the threshold will be converted to zero (0) or black, while pixels with intensities above the threshold will be converted to one (1) or white, resulting in a binary image. This binary image is handy for further segmentation, such as in detecting specific areas, including prostate cancer detection in MRI images.

b) Segmented Mask with ROI (Region of Interest) uses the FindContours technique to detect the boundaries or edges of objects in the binary image that has been obtained through the thresholding process, FindContours works by looking for significant changes in pixel intensity, resulting in contours that can be used to mark areas that are considered the main object, after the contour is found, a binary mask is created to highlight and extract the ROI area. Thus, ROI-based segmentation allows the extraction process to focus more on the relevant part of the image without interference from the insignificant background. Creating a mask to focus on the prostate region with the findcontours method.

3.6. Feature Extraction using Grey Level Co-Occurrence Matrix (GLCM)

At the feature extraction stage, the GLCM method is used to extract texture characteristics from preprocessed and segmented image data.

Feature extraction includes contrast, correlation, energy, and homogeneity features.

Then use four angular directions for feature calculations: 0°, 45°, 90°, and 135°. So that later, the data generated from the feature extraction is tabular, consisting of 16 features, as shown in Table 2.

Table 2. Detail feature extraction

Features	Orientations			
	0°	***45°***	***90°***	***135°***
Contrast	contrast_0	contrart_45	contrast_90	contrast_135
Correlation	correlation_0	*correlation_45*	*correlation_90*	*correlation_135*
Energy	energy_0	energy_45	energy_90	energy_135
Homogeneity	homogeneity_0	energy_45	homogeneity_90	homogeneity_135

3.7. Tabular Data Preprocessing

Preprocessing of the extracted data is done to prepare the data before training the classification model with XGBoost. The preprocessing stages performed are:

a) Label encoding is a preprocessing method that converts labels with categorical data types into numerical values. The numerical class label is 0 for a positive label and 1 for a negative label.
b) Data normalisation is performed using the min-max scaler, ensuring each feature's values are in the range 0 to 1. The purpose of normalising the data on the GLCM feature extraction results is to overcome differences in scale between features, so that the model can learn and train the data without being disturbed by other specific features.

3.8. Data Segmentation

After preprocessing tabular data, the next step is to divide the data into three subsets: training, validation, and test. Data segmentation is performed with 0.8 for training, 0.1 for validation, and 0.1 for testing, using the total data. This division of the composition aligns with research [26] because it shares similarities with the classification algorithm used.

The training data will be used to train the XGBoost algorithm; the validation data will be used to evaluate the model's performance; and the test data will be used to assess the model's performance.

3.9. Data Undersampling

After the data are divided into three subsets, the next step is undersampling to address class imbalance. Undersampling is applied to the training data to reduce dominance or bias from a class with more data during training and to reduce overfitting in the model.

3.10. Model Development using Extreme Gradient Boosting (XGBoost)

The classification model used in this experiment is XGBoost, a boosting method.

This algorithm works by gradually building a decision tree, where each new tree attempts to correct the previous tree's errors by minimising the loss function.

Hyperparameters used in the model include:

- n_estimators = 200 (Number of decision trees)
- learning_rate = 0.05 (Model learning rate)
- max_depth = 8 (Maximum depth)
- subsample = 0.6 (Proportion of data used per iteration)
- colsample_bytree = 0.7 (Proportion of features selected in each tree)
- eval_metric = "logloss" (Evaluation method during training)
- early_stopping_rounds = 20 (Stopping training if there is no improvement)

The training process uses early stopping on validation data to avoid overfitting.

3.11. Evaluation Model using Confusion Matrix

The model has been created and trained, and it is validated on the validation dataset and tested on the test dataset. Furthermore, evaluation is performed using each dataset with the following metrics:

- Accuracy → Measures the percentage of correct predictions.
- Precision & Precision Macro → Measures the accuracy of optimistic predictions.
- Recall & Recall Macro → Measures how well the model detects all positive cases.
- F1-Score & F1-Score Macro → Harmonised average between precision and recall.
- Confusion Matrix → Displays the comparison between the prediction and the original label.

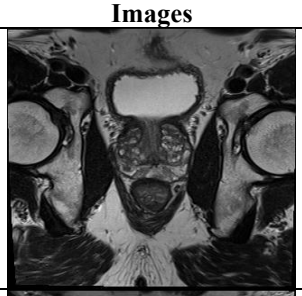
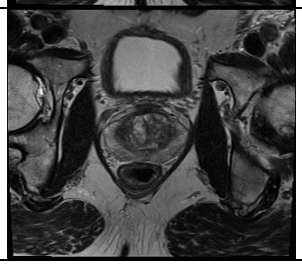
4. Results and Discussion

The outcomes for prostate cancer with XGBoost models are presented in this section. Specifically, this study focuses on the assessment of the Grey Level Co-occurrence Matrix (GLCM) and classification model using the Extreme Gradient Boosting (XGBoost) classifier. These feature extraction and classifiers were examined in detail, with the aim that performance assessment across various training, validation, and test sets will reveal their capabilities and the level of accuracy they achieve.

4.1. Data Collection

Prostate MRI images derived from FastMRI are the dataset used in this experiment [25], which is classified into two categories: Positive (Cancer Detected) and Negative (No Cancer Detected). The total number of images collected is 961, distributed as follows: 424 images for the positive class and 537 images for the negative class, as shown in Table 3.

Table 3. Examples of Image Data

No	Class	Images
1	Positive	
2	Negative	

4.2. Image Data Preprocessing

Figure 3 represents the preprocessed prostate MRI scans, which have undergone several stages before being used for analysis or prostate cancer classification. The images have been resized to 224x225 pixels and converted to grayscale, retaining only pixel intensity information without colour data. Then apply noise-reduction techniques, such as median filtering, and adjust contrast using contrast stretching to enhance contrast, making it easier to distinguish between normal and abnormal areas.

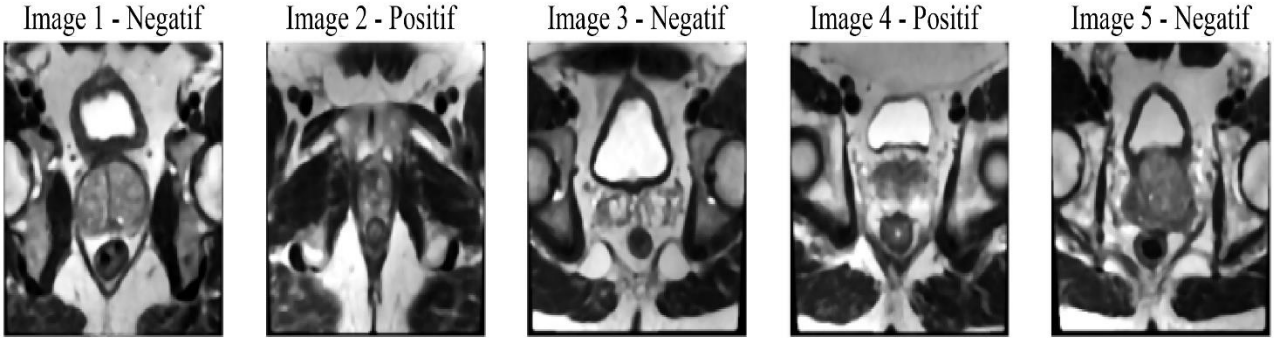


Fig. 3 Result Image Preprocessing

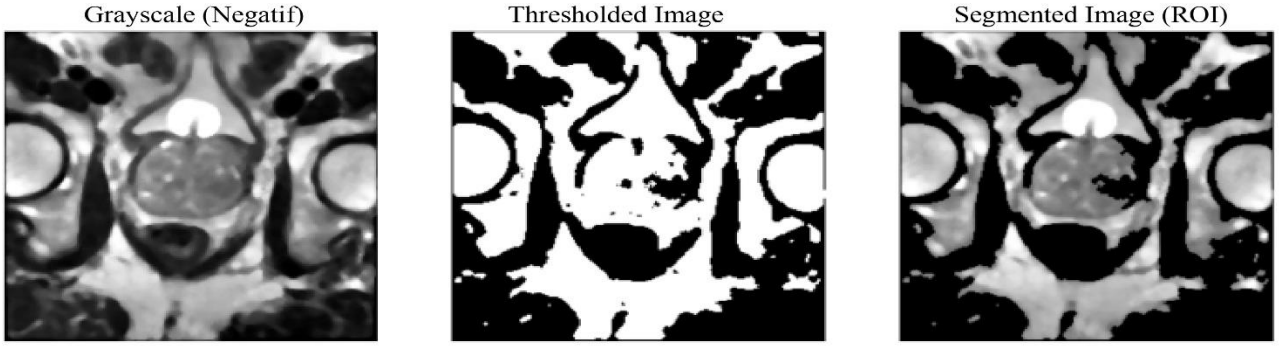


Fig. 4 Result image segmentation

4.3. Image Segmentation

The image illustrates the preprocessing and segmentation stages of a prostate MRI scan. The grayscale image (left) represents the original scan, converted to a single-channel format to preserve intensity variations. The thresholded image (middle) uses Otsu's thresholding to separate the foreground and background based on pixel intensity, enhancing contrast for segmentation. The segmented image (ROI) (right) is generated using contour detection, isolating the prostate gland for analysis. This process enhances feature extraction and improves classification accuracy by reducing noise and background interference, as shown in Figure 4.

4.4. Feature Extraction using Grey Level Co-Occurrence Matrix (GLCM

In this study, the Grey Level Co-Occurrence Matrix (GLCM) was employed to extract texture-based features—contrast, correlation, energy, and homogeneity — from prostate MRI images. These extracted features provide valuable insights into the structural and textural patterns within the pictures. The results of the feature extraction process are presented and analysed to assess their impact on classification performance.

4.5. Tabular Data Preprocessing

The preprocessing of tabular data involved several key steps to ensure optimal model performance. Label encoding and normalisation were applied to preprocess the tabular data, ensuring all features were in a uniform numerical format and scale before breaking the data into three parts: training, validation, and testing. Random undersampling was then used to balance the classes to 424 samples each. These preprocessing steps helped improve data quality and ensure an effective learning process for the classification model.

4.6. XGBoost Model

In this section, we will discuss the results of the Boosting classification model, namely XGBoost. XGBoost, as a boosting-based ensemble learning method, has the advantage of handling complex data and improving accuracy through stepwise modelling—analysis through the learning curve to evaluation. The results obtained will be compared to identify potential overfitting, underfitting, and model fit. The logloss curve (Figure 5) presents the training and validation log-loss over multiple iterations (up to 200). Key observations include:

a) Training Logloss: The blue line shows a steady decline, indicating that the model is learning and optimising its predictions over iterations.
b) Validation Logloss: The orange line decreases at the beginning, then stabilises, suggesting that further iterations may not significantly improve generalisation performance.

Analysis of the confusion matrix for each data subset (training, testing, and validation) was used to monitor the performance of the XGBoost model. By comparing the results across each data subset, it can be seen whether the model is overfit, underfit, or already generalises well to new data. The following discusses each of the Confusion Matrices obtained.

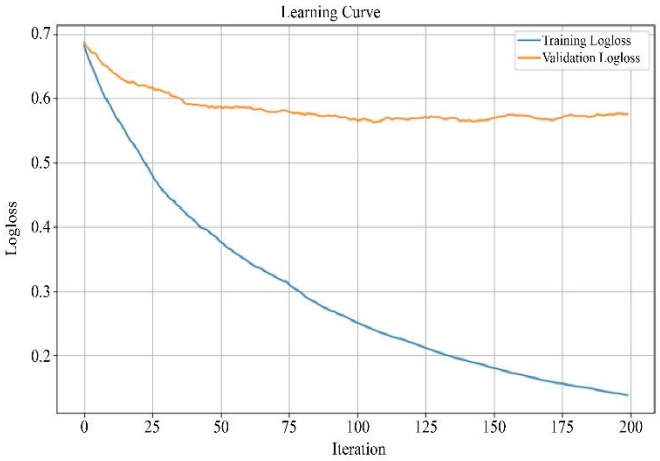


Fig. 5 The Training and Validation Curve Logloss

4.6.1. Confusion Matrix for Training Set

In Figure 6, the models on the training data perform exceptionally well, with 339 True Negatives (TN) and 338 True Positives (TP). There are also zero False Negatives (FN) and only one False Positive (FP), indicating almost perfect Classification on the training set. This result from the confusion matrix suggests that the model has learned the patterns in the training data very effectively, but it may also indicate a risk of overfitting.

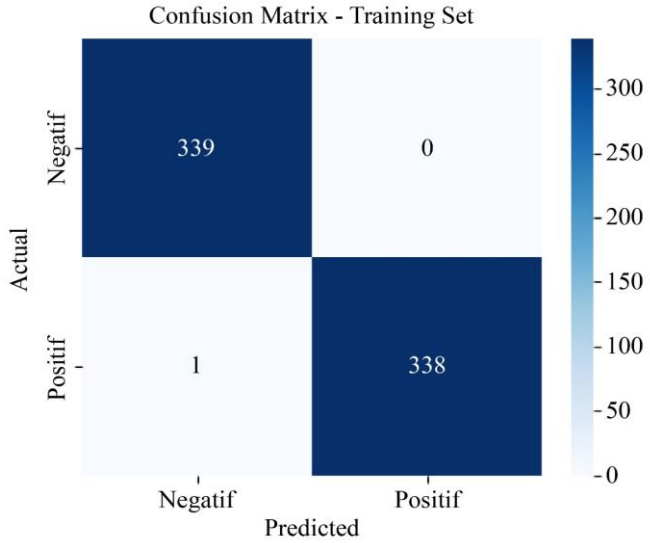


Fig. 6 Confusion Matrix for Training Set

$$Accuracy = \frac{(338+339)}{(338+339+0+1)} = 0.9985 \; or \; 99{,}85\%$$

$$Precision = \frac{338}{(338+0)} = 1 \; or \; 100\%$$

$$Recall = \frac{338}{(338+1)} = 0.9970 \; or \; 99{,}7\%$$

4.6.2. Confusion Matrix for Validation Set

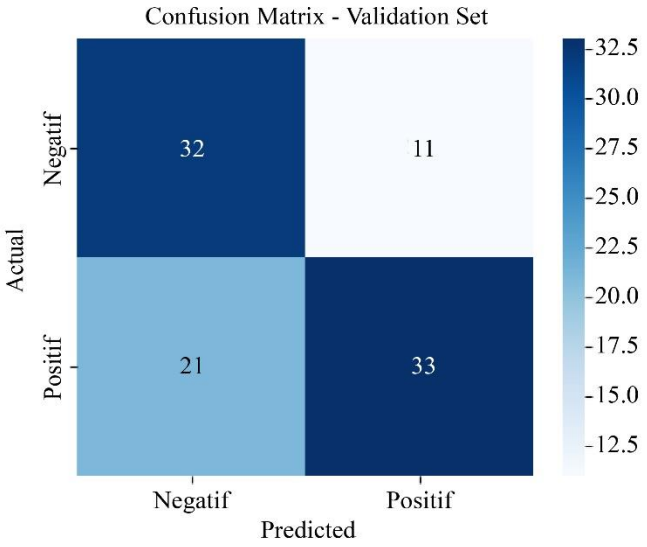


Fig. 7 Confusion matrix for validation set

In Figure 7, the model's performance drops on the validation set, indicating the model struggles more with unseen data. The model correctly classifies 32 True Negative samples (TN) and 33 True Positive samples (TP). However, it misclassifies 11 negative samples as False Positives (FP) and 21 positive samples as False Negatives (FN). The increased number of misclassifications suggests that the model does not generalise as well as expected to the validation set.

$$Accuracy = \frac{(33+32)}{(33+32+21+11)} = 0.6321 \; or \; 63{,}21\%$$

$$Precision = \frac{33}{(33+11)} = 0.75 \; or \; 75\%$$

$$Recall = \frac{33}{(33+21)} = 0.6111 \; or \; 61{,}1\%$$

4.6.3. Confusion Matrix for Testing Set

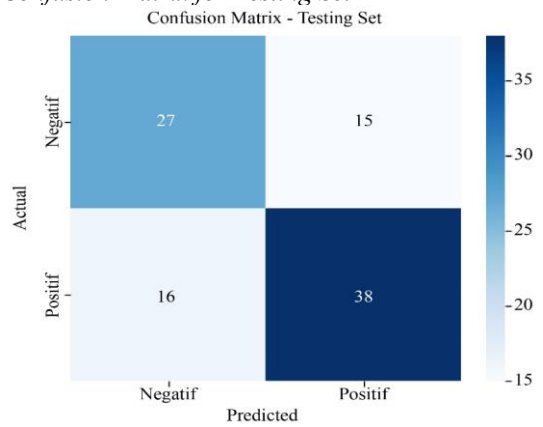


Fig. 8 Confusion Matrix for Testing Set

In Figure 8, the model's performance is still not optimal on the test set, with correctly classifying 27 True Negative samples (TN) and 38 True Positive samples (TP). Numbers 15 FP and 16 FN indicate that the model is making a considerable number of errors. The balanced number of false positives and false negatives suggests that the model has difficulty effectively distinguishing between the two classes on unseen data.

$$Accuracy = \frac{(38+27)}{(38+27+15+16)} = 0.6321 \; or \; 63{,}21\%$$

$$Precision = \frac{38}{(38+15)} = 0.7169 \; or \; 71{,}69\%$$

$$Recall = \frac{38}{(38+16)} = 0.7037 \; or \; 70{,}37\%$$

Table 4. Model Evaluation

Dataset	Accuracy	Precision	Recall
Training	99.8%	100%	99.7%
Validation	63.2%	75%	61,1%
Testing	63.2	71.6%	70.37

Table 4 shows the evaluation results, with the model performing well during training, achieving an accuracy of 99.8%, demonstrating a strong ability to learn discriminative features. Accuracy on the validation and test datasets yielded stable generalisation performance on the evaluation data, demonstrating its ability to recognise patterns.

The model effectively captured complex feature relationships despite variations in training and testing performance. The precision on the testing dataset (71.6%) and the recall (70.37%) demonstrated balanced classification performance for the model used.

5. Conclusion

Prostate cancer classification using MRI images using the Extreme Gradient Boosting (XGBoost) model has been successfully performed using Grey Level Co-occurrence Matrix (GLCM) feature extraction. The texture-based feature identification results differentiate cancerous and non-cancerous tissue through a hybrid approach that provides an appropriate framework. This model effectively learns discriminatory patterns, with generalisation maintained through 99.8% accuracy on the training data and 63.2% on the validation and test datasets, respectively. The model also achieved balanced classification performance, potentially aiding medical diagnosis, with precision and recall values exceeding 70% on the testing dataset. Near-perfect training accuracy indicates that general patterns have been learned, but there is a tendency for overfitting. The model is not yet fully robust on healthcare datasets, requiring hyperparameter tuning for further research. Augmentation and regularisation techniques should also be explored to produce a model that fits and generalises across diverse, large datasets.

Prostate cancer detection from MRI images is well-suited to the GLCM approach with the XGBoost model. This research has contributed to medical images that serve as references demonstrating that Machine learning is practical for early cancer detection, decision support systems, and clinical image analysis.

Acknowledgments

We want to express our deepest gratitude to Universitas Pembangunan Nasional Veteran Jakarta for the support, facilities provided for this research, and financial assistance, which enabled this research to be carried out. We would also like to thank Universiti Teknikal Malaysia Melaka and RSPAD Hospital for their support for this research. We would also like to express our appreciation to our colleagues for their discussions and support. Finally, we would like to thank our supervisors at RSPAD Hospital. This research is not biased or has any interests in any institution; it is purely intended to develop this research so that it can be helpful to the general public.

Turnitin Similarity and AI Verification

The manuscript has been verified using Turnitin. The similarity index is 4%, and the AI-generated content detection result is 0%, ensuring originality and compliance with IJETT publication standards.

Data Availability

The data that support the findings of this study are openly available in [zenodo] at https://zenodo.org/records/16910415

Ethical Consideration

This research utilised anonymised and publicly available MRI datasets. Hence, no personal data was used, and no ethical approval or patient consent was required.

Credit Author Statement

Neny Rosmawarni: Research leader, Conceptualisation, Data Collection, Methodology, Supervision, Writing – Review & Editing. Zeratul Izza Mohd Yusoh: Resources, Data Curation, Methodology, Formal Analysis, Writing – Review & Editing. Yun Houy Choo: Validation, Formal Analysis, Writing – Review & Editing. Thoyyibah T: Conceptualisation and Design, Processing Data, Formal Analysis, Writing – Original Draft, Supervision. Karunia Agustiani: Processing Data, Visualisation, Writing – Original Draft. Nadra: Investigation, Interpretation, Resources, Writing – Review & Editing.

References

[1] Sobia Wasim, Sang-Yoon Lee, and Jaehong Kim, "Complexities of Prostate Cancer," *International Journal of Molecular Science*, vol. 23, no. 22, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Eka Yudha Rahman et al., "Empowering Health Workers in Early Detection of Prostate Cancer at the North Banjarbaru Community Health Center," *ILUNG Community Service Journal (Superior Wetland Innovation)*, vol. 3, no. 3, pp. 516-526, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fari Katul Fikriah, Amelia Devi Putri Ariyanto, and Arif Fitra Setyawan, "Classification of MRI Results of Brain Tumors with GRAY Level Co-Occurance Matrix (GLCM) Feature Extraction," *Rabit Journal of Technology and Information Systems Univrab*, vol. 9, no. 2, pp. 343-350, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Rafli Dika Pramudya, "*Classification of Brain Tumors based on MRI Images with the Random Forest Classifier Method using GLCM Feature Extraction*," Thesis, Veteran National Development University Jakarta, 2024. [Google Scholar] [Publisher Link]

[5] Adi Muzakir, Anita Desiani, and Ali Amran, "Classification of Prostate Cancer Disease Using Naïve Bayes and K-Nearest Neighbor Algorithms," *Komputika: Journal of Computer Systems,* vol. 12, no. 1, pp. 73-79, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Acep Handika Permana, Fajri Rakhmat Umbara, and Hope Kasyidi, "Classification of Cardiovascular Type Heart Disease using Adaptive Synthetic Sampling and Extreme Gradient Boosting Algorithms," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 1, pp. 499-508, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Kairupan Indah Yessi et al., "An Extreme Gradient Boosting Approach for Classification and Sentiment Analysis," *The Asian Journal of Technology Management*, vol. 16, no. 3, pp. 211-225, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Hasria Alang Hafsah, and Muh Sri Yusal, "Increasing Knowledge of 'Prostate Cancer' in the Community of Mapung Buttu Hamlet, Campalagian District, Polewali Mandar Regency," *CREATIVE: Indonesian Journal of Community Service*, vol. 2, no. 2, pp. 1-6, 2022. [Google Scholar]

[9] Viva Ratih Bening Ati, "Risk Factors Associated with the Incidence of Prostate Cancer (Case Study at Prof. Dr. Margo Soekarno Purwokerto Regional Hospital)," *Mandala of Health*, vol. 14, no. 2, pp. 67-73, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hanisa Hanisa, I Putu Eka Juliantara, and Edwien Setiawan Saputra, "Cervical MRI Examination Management in Cases of Spinal Cord Syringomyelia at Primaya Hospital, Tangerang," *Journal of Research in the Medical Science Cluster*, vol. 2, no. 2, pp. 30-40, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rio Aditya Pahlevi, and Bayu Setiaji, "Analysis of Application Haar Cascade Classifier and Local Binary Pattern Histogram Algorithm in Recognizing Faces with Real-Time Grayscale Images using Opencv," *Journal of Informatics Engineering (Jutif)*, vol. 4, no. 1, pp. 179-186, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Orry Adrianus Mokola, "Predicting Honey Production Amount Based on Honeycomb Size Using Image Processing Method," *Animal Science Research Journal*, vol. 1, no. 2, pp. 31-41, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Hotma Pangaribuan, and Sunarsan Sitohang, "Improving Edge Detection Quality with Image Segmentation Method," *Remik*, vol. 7, no. 1, pp. 591-601, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Muhammad Fakhrurrozy Cahyadi, Siswan Syahputra, and Mili Alfhi Syari, "Implementation of Thresholding Method in Digital Image Transformation Process," *Educate Journal of Educational Sciences and Teaching*, vol. 1, no. 3, pp. 319-346, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Kartika Candra Kirana et al., "Classification of Brain Tumor Disease using K-Nearest Neighbor based on GLCM," *Tekno Journal of Electrical and Vocational Technology*, vol. 33, no. 1, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Epimack Michael et al., "Breast Cancer Segmentation Methods: Current Status and Future Potentials," *BioMed Research International*, vol. 2021, no. 1, pp. 1-29, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Desi Nurnaningsih et al., "Identification of Rhizome Type Medicinal Plant Images Using Euclidean Distance Based on Shape and Texture Characteristics," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 3, pp. 171-178, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Adrian Budi Prawira, Jayanta Jayanta, and Yuni Widiastiwi, "Application of the Gray Level Co-Occurance Matrix Method and the Support Vector Machine Algorithm in Classifying Jujube Plants Based on Leaf Texture," *Proceedings of the National Seminar on Computer Science and its Applications*, vol. 2, no. 1, pp. 569-578, 2021. [Google Scholar] [Publisher Link]

[19] Lei Huang et al., "Normalization Techniques in Training DNNs: Methodology, Analysis and Application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10173-10196, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto, "Comparison of Min-Max Normalization and Z-Score Normalization in the K-Nearest Neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *International Journal of Informatics and Information Systems*, vol. 4, no. 1, pp. 13-20, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Pattaramon Vuttipittayamongkol, Eyad Elyan, and Andrei Petrovski, "On the Class Overlap Problem in Imbalanced Data Classification," Knowledge-Based Systems, vol. 212, pp. 1-55, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Wilda Imama Sabilla, and Candra Bella Vista, "Implementation of SMOTE and under Sampling on Imbalanced Datasets for Predicting Company Bankruptcy," *Journal of Applied Computing*, vol. 7, no. 2, pp. 329-339, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Jan Melvin Ayu Soraya Dachi, and Pardomuan Sitompul, "Comparative Analysis of the XGBoost Algorithm and the Random Forest Ensemble Learning Algorithm on Credit Decision Classification," *Research Journal of Mathematics and Natural Sciences Cluster*, vol. 2, no. 2, pp. 87-103, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Joan Y. Chiao, *Philosophy of Computational Cultural Neuroscience*, Machine Learning, 1st ed., New York, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Radhika Tibrewala et al., "FastMRI Prostate: A Public, Biparametric MRI Dataset to Advance Machine Learning for Prostate Cancer Imaging," *Scientific Data*, vol. 11, no. 1, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Pratima Kumari, and Durga Toshniwal, "Extreme Gradient Boosting and Deep Neural Network based Ensemble Learning Approach to Forecast Hourly Solar Irradiance," *Journal of Cleaner Production*, vol. 279, 2021. [CrossRef] [Google Scholar] [Publisher Link]