MCA Major Project report on

Pancreatic Tumor Detection using Image Processing and Machine Learning.

Project Submitted in Partial Fulfilment of the degree of

Master of Computer Applications

Submitted By:

Gaurav Singh Rawat,

Examination Roll No: 21234757020

Year: 2025

Submitted To:

Dr. Mantosh Biswas

Associate Professor

Department of Computer Science,

University of Delhi

UNDERTAKING

I declare that the work presented in this report titled “Pancreatic Tumor Detection” submitted to the Department of Computer Science, University of Delhi, New Delhi, for the award of the Master of Computer Application degree, is my original work. I have not plagiarized or submitted the same in part or full to any university or institution for the award of any degree or diploma. In case this undertaking is found incorrect, I accept that our degree may be unconditionally withdrawn.

Signature of Candidate

Date: 16/07/2025

Gaurav Singh Rawat

(Exam Roll. No. 21234757020)

Place: Delhi

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF DELHI, DELHI – 110 007
(INDIA)

CERTIFICATE

I hereby certify that the work which is being presented in this MCA Major Project report entitled “Pancreatic Tumor Detection using Image processing and ML”, in partial fulfilment of the requirements for the award of the Master of Computer Application is an authentic record of my own work carried out during a period from January, 2025 to June, 2025 under the supervision of mentor Dr. Mantosh Biswas, Associate Professor, Department of Computer Science, University of Delhi, The matter presented in this project report has not been submitted for the award of any other degree elsewhere.

Date: 16/07/2025

Place: Delhi

Signature of Candidate
Gaurav Singh Rawat
(Exam Roll No.21234757020)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Signature of Faculty Mentor	Signature of HoD
Associate Professor	Professor

Acknowledgement

I would like to take this opportunity to sincerely thank everyone who supported and guided me throughout the completion of this project. my deepest thanks go to **Dr. Mantosh Biswas**, Associate Professor, Department of Computer Science, University of Delhi, for his valuable mentorship, clear direction, and continuous encouragement throughout this work. His support helped me stay focused and motivated from start to finish. I also extend my gratitude to **Mr. Rajesh Kumar Yadav**, Ph.D. Scholar, Department of Computer Science, for his technical guidance, helpful insights, and readiness to assist whenever needed. His suggestions greatly improved the quality and clarity of the project. I would like to place on record my deep sense of gratitude to **Professor Neelima Gupta, Head of Department** of the Department of Computer Science, University of Delhi for providing this opportunity. Lastly, I would like to thank my **family and friends** for their constant encouragement, patience, and moral support. Their belief in me helped me remain determined and confident during all phases of this project.

Gaurav Singh Rawat

21234757020

Index

Table1 of Abbreviations: - Short form used in this project.

Abbreviation	Full Form
CT	Computed Tomography
SVM	Support Vector Machine
RBF	Radial Basis Function
GLCM	Gray Level Co-occurrence Matrix
PCA	Principal Component Analysis
SMOTE	Synthetic Minority Over-sampling Technique
FCC	Freeman Chain Code
MCC	Matthews Correlation Coefficient
AUC	Area Under the Curve
ROI	Region of Interest
CPU	Central Processing Unit
GPU	Graphics Processing Unit
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CV	Cross Validation
Abbreviation	**Full Form**
Joblib	Job Library (for saving/loading Python objects like models and scalers)
XGBoost	Extreme Gradient Boosting
VS Code	Visual Studio Code

CLAHE	Contrast Limited Adaptive Histogram Equalization
OpenCV	Open-Source Computer Vision Library
.npy	NumPy File Format (used for saving NumPy arrays)
Sklearn	scikit-learn (Python ML library)
Pywt	PyWavelets (Python Wavelet Transform library)
CAD	Computer-Aided Diagnosis
LBP	Local Binary Pattern
DWT	Discrete Wavelet Transform
CNN	Convolutional Neural Network

1. Introduction

A pancreatic tumor is a lump or abnormal growth that forms inside the pancreas. The pancreas is an important organ that helps in digestion and in controlling blood sugar. Some tumors are harmless (called benign), but others can be dangerous (called malignant or cancerous). Pancreatic cancer is one of the most serious types of cancer because it grows quickly and is often not found until it is too late. That is why it is very important to find it as early as possible and for that we use the images CT scan mostly used by the doctors to find tumors in early stages. CT stands for Computed Tomography. A CT scan is a special type of X-ray machine that takes many pictures of the inside of the body. These pictures are then combined to create detailed images of internal organs. In this project, we use CT scan images of the pancreas to check if a tumor is present. These images are in black and white (grayscale) and show the shapes and structures inside the body. Which makes them very good for Early Detection of tumor. Pancreatic cancer is usually not found until it is in the final stage. At that point, treatment becomes very difficult and the chances of survival are low. If we can detect the tumor early, doctors can begin treatment sooner, and the patient has a better chance of recovery. But manually checking hundreds or thousands of CT images takes a lot of time and effort. So, using Machine Learning to do this job can help doctors save time and avoid mistakes. ML is a way to teach computers to make decisions by learning from data. Just like humans learn by looking at examples, computers can learn by looking at many images that are already labelled (like "tumor" or "normal"). After learning from these examples, the machine can look at new images and decide whether they show a tumor or not. In this project, we use machine learning to build a system that can identify tumors in CT scan images. These tumor and non- tumor are called labels which are used by supervised machine learning algorithm.

SVM is a supervised machine learning algorithm used for classification. SVM tries to find the best possible boundary (hyperplane) that separates classes with the maximum margin. In your tumor detection project, it is used to classify images as either normal or tumor. If the data is simple and clearly separable, SVM can draw a straight line between the two groups. But medical image data is often complex and not easy to separate with a straight line. This is where we use kernel function in SVM to handle such cases. A kernel helps the SVM convert the data into a higher dimensional space where it becomes easier to draw a clear boundary between the two classes. For this project, we use the RBF (Radial Basis Function) kernel, which is very good at handling curved and complex patterns in image data. We chose SVM for this project because it works well with small to medium-sized datasets, it gives very good results even when the data is complex, it can separate images into tumor and normal groups very clearly, it does not get confused easily, even if there are some errors in the data and it works well when combined with other techniques like PCA (to reduce extra information) and SMOTE (to balance the data). In many hospitals, doctors look at CT images one by one to check for tumors. This takes a lot of time and effort. Sometimes, small tumors can be missed. This project creates an automatic system that can quickly and correctly detect pancreatic tumors from CT images. It helps doctors by giving them fast and accurate results so that they can make better decisions for patients. The main goal of this project is to build a smart computer program that can Read and process CT images, Improve the image quality using image processing, Pick out important information (features) from each image, Use PCA to reduce the number of features and keep only the useful ones, Use SMOTE to balance the number of tumor and non-tumor samples, Train a model using SVM and XGBoost to classify the images and Show results like accuracy, sensitivity, specificity, and more to prove how well the system works.

The contribution of research work as follows:

1. The model employs contrast enhancement through CLAHE and denoising techniques to improve the quality of grayscale pancreatic CT images, thereby ensuring more accurate feature extraction and subsequent classification performance.
2. This work introduces a multi-level feature extraction strategy that combines shape-based features, texture analysis using GLCM at multiple distances and angles, and frequency-based features via multi-level wavelet transforms capturing comprehensive spatial and frequency domain information critical for tumor detection.
3. The model addresses the skewed distribution in medical datasets using SMOTE (Synthetic Minority Over-sampling Technique), which synthetically balances the dataset and prevents the model from becoming biased toward the majority class, ultimately improving sensitivity and generalization.
4. A custom ensemble approach combining RBF-kernel SVM (with pairwise_kernels) and XGBoost was used for classification. This combination leverages the strengths of both algorithms—SVM's decision boundaries and XGBoost's boosting power—to deliver higher accuracy, AUC, F1-score, and MCC.
5. The complete pipeline is designed with runtime efficiency using PCA for dimensionality reduction and joblib for saving trained models. This makes the system scalable for real-world deployment in clinical settings, potentially aiding radiologists in early tumor detection.

2. Literature survey

In recent years, tumor detection has gained substantial attention due to its critical role in early diagnosis and improving treatment outcomes. Numerous studies have investigated diverse approaches combining image processing, radiomics, and machine learning to automate and enhance the accuracy of tumor classification. A hybrid model integrating Principal Component Analysis (PCA) with Histogram-based Features and Support Vector Machine-Random Forest (HFB-SVM-RF) demonstrated significant improvement in tumor classification performance by reducing feature dimensionality and utilizing ensemble learning techniques [1]. Convolutional Neural Networks (CNNs) have also shown strong performance in tumor segmentation and classification tasks, particularly in glioma detection through MRI scans, where spatial hierarchies learned by deep layers enhanced diagnostic precision [2]. Multimodal MRI analysis incorporating Support Vector Machines (SVM) has been utilized to fuse structural and functional features such as shape and texture, offering better generalization across heterogeneous tumor presentations [3]. Reviews have further highlighted the widespread applicability of deep learning across medical imaging, emphasizing its capabilities in feature representation learning and end-to-end classification pipelines [4]. Texture and wavelet-based methods have proven useful, as wavelet transforms capture multi-scale frequency details, and when combined with SVM classifiers, these techniques yield high sensitivity in breast cancer classification [5].

Machine learning has also been applied to develop computer-aided diagnostic (CAD) systems that facilitate decision support in medical imaging workflows, particularly through CT and MRI scans [6]. Deep CNN architectures tailored for tumor segmentation have made considerable strides in delineating complex tumor shapes and boundaries in brain MRI datasets [7]. Radiomics, a burgeoning domain, leverages quantitative imaging features extracted from medical scans, and when combined with ensemble classifiers like XGBoost, has been particularly effective for pancreatic tumor detection [8]. Addressing the common challenge of class imbalance in medical datasets, methods such as the Synthetic Minority Oversampling Technique (SMOTE) have shown marked improvements in model performance across rare tumor classes [9]. Transfer learning has emerged as a valuable technique for scenarios with limited labelled medical data, allowing pre-trained deep learning models to be adapted effectively for tumor detection tasks [10].

One noteworthy effort in this direction involves the use of radiomics features derived from endoscopic ultrasound images for pancreatic cancer detection. By extracting Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrix (GLCM) features and employing classifiers like SVM and Random Forest, automated CAD systems have significantly outperformed human interpretation in both sensitivity and specificity [11]. Traditional preprocessing steps such as boundary enhancement and contrast optimization remain essential for successful segmentation and feature extraction. These pre-processed images are labeled and passed through a carefully designed feature extraction pipeline, capturing shape (e.g., boundary descriptors), texture (e.g., GLCM, statistical moments), and frequency-domain features (e.g., wavelet decomposition) for comprehensive image representation [12][13]. These features are then combined into a unified feature vector, scaled for normalization, and further reduced in dimensionality using PCA to improve training efficiency and model generalizability [14]. Class balancing via SMOTE ensures fair representation during model learning. Models are then trained using robust classifiers, including SVM with custom kernels and ensemble methods like Bagging and XGBoost. Performance is assessed using key metrics such as sensitivity, specificity, F1 score, and AUC derived from ROC analysis.

Recent studies also addressed deep learning alternatives in pancreatic tumor detection. For example, a method developed for analysing full video sequences of endoscopic ultrasound (EUS) examinations demonstrated high specificity (>90%) without pre-selection of suspicious frames or reliance on deep neural networks. This strategy proved robust even when implemented in non-optimized MATLAB code, processing frames within 0.465 seconds on standard hardware [15]. Other approaches used elastography and contrast-enhanced imaging in conjunction with multilayer perceptrons, though their effectiveness is limited by reliance on intravenous contrast and variability in manual procedures [16][17].

Fusion-based models have also gained traction. By combining clinical features with imaging-based radiomics, studies demonstrated that models incorporating both domains achieve superior performance. For instance, a model combining 28 clinical parameters with 306 radiomics features achieved an AUC of 0.978 in the training set and 0.925 in the test set for pancreatic tumor classification [18][19]. Deep learning frameworks built on U-Net and its derivatives have also been employed to detect pancreatic ductal adenocarcinoma (PDAC) in contrast-enhanced CT scans. These models used anatomical context and voxel-level tumor likelihood mapping for robust early-stage tumor detection, particularly those <2cm, which are typically difficult to diagnose [20].

Novel object detection architectures such as augmented Feature Pyramid Networks (FPN) combined with Faster R-CNN have been shown to outperform state-of-the-art detectors, achieving 90.18% accuracy and an AUC of 0.9455 in pancreatic tumor classification. The integration of bottom-up feature propagation, self-adaptive fusion, and contextual dependency Modeling helped capture tumors of varying sizes and shapes more effectively [21]. Shape and texture fusion strategies were also explored in FusionNet, where separate branches processed binary segmentation masks and raw CT images respectively. These fused representations led to remarkable performance improvements with sensitivity of 92.65% and specificity of 97%, particularly beneficial in PDAC cases where lesions subtly alter the pancreas structure [22].

Approaches incorporating clinically-relevant secondary anatomical features—such as the pancreatic duct and bile duct—into tumor detection pipelines were also found to be effective. A model integrating such information into a 3D U-Net achieved 99% sensitivity and specificity, showing how mimicking radiologist strategies can lead to both interpretable and accurate computer-aided diagnoses [23]. Non-imaging biomarkers such as thermal liquid biopsy (TLB) were also explored. A machine learning model analysing serum thermograms from PDAC patients achieved high accuracy and AUC scores by focusing on proteomic transitions in specific temperature ranges, offering a minimally invasive diagnostic path [8].

These diverse studies collectively highlight the increasing sophistication of machine learning models in tumor detection, particularly pancreatic tumors. The integration of radiomics, clinical data, and model architectures continues to drive improvements in diagnostic accuracy, model interpretability, and computational efficiency. While deep learning dominates many recent contributions, hybrid approaches combining traditional feature engineering with modern ensemble and transfer learning methods still prove to be highly effective, especially in data-scarce medical scenarios.

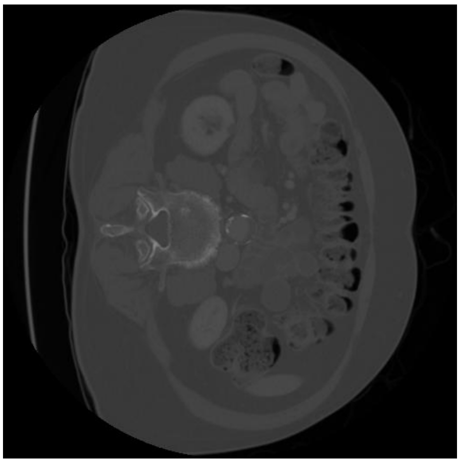
3. Proposed Methodology

The pancreatic tumor detection begins with the acquisition of medical images and then we do Preprocessing to improve visual quality and enhance relevant features like contrast and boundaries. Through Labelling each image is assigned a class label to distinguish between tumor and non-tumor cases. After preprocessing, we do Feature Extraction where meaningful features are extracted to represent the shape, texture, and frequency characteristics of the image. These may include morphological descriptors, texture patterns, and wavelet-based information to capture both spatial and frequency details. Then through Feature combination extracted features from all categories are then merged into a single comprehensive feature vector for each image. We use Feature Scaling to ensure consistent scaling across different feature types, normalization or standardization techniques are applied. Dimensionality reduction techniques such as PCA are introduced to reduce the complexity of the feature space while retaining significant information. Dataset Balancing technique is used to address class imbalance, a data balancing strategy like oversampling or SMOTE is used. The processed dataset is then divided into training and testing sets for model development aka Train-Test split Technique. Through Model Training a machine learning model is trained using classifiers possibly with custom kernel functions or ensemble methods such as bagging to improve generalization and accuracy. Finally, the model is used to make predictions on unseen data, and its performance is evaluated using various classification metrics and visualization tools such as ROC curves and Precision-Recall Technique.

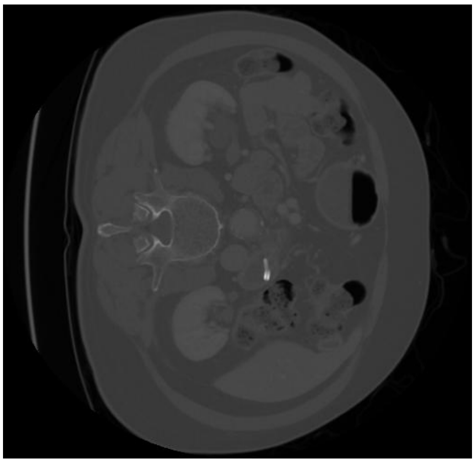
The experimental setup for this project relies on a variety of tools and technologies, each playing a specific role in the overall pipeline. The primary programming language employed is Python due to its simplicity, readability, and extensive ecosystem for scientific computing and machine learning. For image-related tasks such as loading, resizing, enhancement, and preprocessing, a robust computer vision library is used, which allows efficient manipulation of medical images. Numerical computations and array operations are handled by a high-performance scientific computing library, which supports fast and efficient matrix manipulations crucial for feature extraction. Machine learning models, data preprocessing techniques, and evaluation metrics are implemented using a comprehensive machine learning library that offers a wide range of tools for classification, scaling, dimensionality reduction, and performance assessment. A specialized and powerful gradient boosting framework is utilized for classification tasks to improve prediction accuracy. To address the issue of class imbalance in the dataset, a synthetic sampling technique is applied to generate new data points for the minority class, ensuring balanced training. For extracting multi-resolution frequency-based features, a dedicated wavelet transform library is used, which facilitates the decomposition of images into various sub-bands. Lastly, to store trained models and reusable processing objects like dimensionality reducers and scalers, a lightweight and efficient object serialization library is adopted, enabling easy reuse and deployment of the trained pipeline.

3.1 Image Loading and Preprocessing

The process begins with the loading of grayscale medical images from two separate directories. one for normal pancreatic images and the other for pancreatic tumor images. This is handled using the load_images () function, where each image is read using OpenCV's cv2.imread with the flag cv2.IMREAD_GRAYSCALE, converting them to a single-channel grayscale format for simplicity and focus on texture patterns. Once loaded, all images are resized to a fixed dimension of 128×128 pixels using cv2.resize to ensure uniform input shape for further processing and model compatibility. A critical step in medical image preprocessing is contrast enhancement, achieved through CLAHE (Contrast Limited Adaptive Histogram Equalization). This is implemented via OpenCV's cv2.createCLAHE, configured with a clip limit of 2.0 and a tile grid size of 8×8, and then applied to each image using clahe.apply(img). CLAHE helps in enhancing local contrast, making subtle differences in intensity more visible, especially in regions of interest such as tumors. This preprocessing ensures that image data is clean, uniform, and well-prepared for meaningful feature extraction.



Normal (non-tumor) CT Scan image



Pancreatic Tumor CT Scan image

Figure 1: -(a) image of Normal and (b)Pancreatic Tumor CT scan.

3.2 Feature Extraction:

Once pre-processed, images are analysed using three core feature extraction techniques which is following.

1. Shape Feature Extraction

To capture the geometric and structural characteristics of tumor regions in the CT images, shape features are extracted from the pre-processed images. First, the images are converted into a simplified form where the tumor region stands out distinctly from the background. This involves highlighting the object of interest (suspected tumor) while suppressing background information. To ensure clarity and reduce unwanted speckles or small gaps, an image cleaning technique is applied that smooths out irregularities and fills small holes, producing a more refined outline of the tumor region. Once the image is clean and the tumor boundary is well-defined, the outermost edge of the tumor is traced. If there are multiple such boundaries detected, the one covering the largest area is selected, as it likely represents the actual tumor. From this boundary, three important shape-related measurements are derived. The first is the total area enclosed by the boundary, indicating the size of the tumor. The second is the perimeter, representing the total length around the boundary. The third is a more advanced measurement called the Freeman Chain Code, which captures the directional sequence along the tumour's outline. This sequence reflects how jagged or smooth the tumor boundary is, offering a deeper insight into its shape complexity. These combined features help characterize tumors more effectively for classification.

Circularity (Shape Feature)

$$Circularity = \frac{4\pi \cdot Area}{Perimeter^2} \qquad (1)$$

2.Texture Feature Extraction (GLCM)

Texture analysis plays a crucial role in medical image processing by enabling the identification of differences between healthy and abnormal tissues based on spatial patterns. Tumor regions tend to exhibit distinct texture characteristics when compared to normal areas, which can be quantified using statistical techniques. One widely used method for extracting such texture features is the Gray-Level Co-occurrence Matrix (GLCM), which analyzes how frequently pairs of pixel values occur in a specific spatial arrangement. In this process, the relationship between horizontally adjacent pixels at a distance of one pixel is considered, using 256 grayscale levels. The resulting matrix is normalized and made symmetric to ensure uniformity across all images. From this matrix, three significant texture descriptors are computed: contrast, which captures intensity variation between neighboring pixels and is often elevated in tumor regions; homogeneity, which reflects the uniformity of adjacent pixel values and is generally higher in normal tissue; and energy, which measures textural orderliness and tends to vary between healthy and tumorous areas. These features collectively help in enhancing the differentiation between normal and tumor CT scans.

1. Contrast (GLCM):

$$\text{Contrast} = \sum_{i,j} |i-j|^2 . P(i,j) \qquad (2)$$

Here i and j are grey levels (intensities) in the image. Where the square of the difference between the gray levels measures how different the neighbouring pixels are. And $.P(i,j)$ function explain the normalized co-occurrence probability between gray levels. Its purpose is to measure local variations in the image. A higher contrast value indicates more intense changes in pixel values.

2. Correlation (GLCM):

$$Correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)\cdot P(i,j)}{\sigma_i \cdot \sigma_j} \qquad (3)$$

μ_i, μ_j mean gray levels along rows and columns of the GLCM. And $\sigma_i \cdot \sigma_j$ Standard deviations of the gray levels along rows and columns. Its purpose is to measures how correlated a pixel is to its neighbour over the whole image. Higher correlation means similar gray levels are often neighbours.

3. Energy (GLCM)

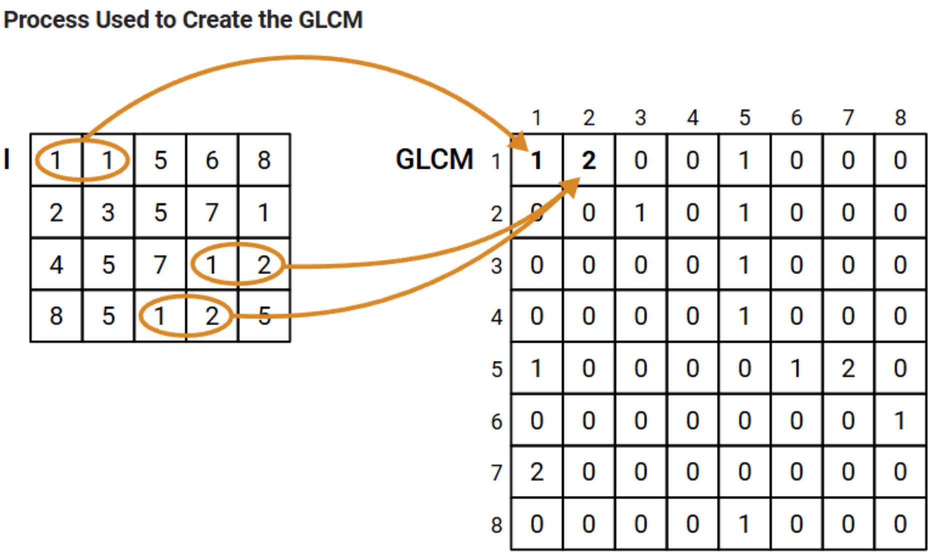
$$Energy = \sum_{i,j} P(i.j)^2 \qquad (4)$$

Here its purpose is to Measure uniformity or texture smoothness. High energy indicates fewer dominant gray-level transitions (more homogenous texture).

4. Homogeneity (GLCM)

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

- Here its purpose is to Measure the closeness of distribution of elements in the GLCM to its diagonal. High value when neighboring pixels have similar intensities.



1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

Figure 2: Image of GLCM Process

2. Frequency Feature Extraction (Wavelet)

Wavelet-based features are used to analyze the frequency content of medical images at multiple resolutions. This approach helps in capturing both fine and coarse details from the CT scans, making it easier to detect subtle variations often associated with tumors. To achieve this, each image is broken down into different components that represent various levels of detail. The process involves a two-level transformation that separates the image into an approximation part (representing low-frequency information) and three detail parts (capturing horizontal, vertical, and diagonal high-frequency information).

From each of these components, statistical values such as the average intensity and the variation in intensity are measured. These values describe how image brightness changes across different regions and scales. Tumors can alter these frequency patterns in unique ways, making such features highly effective for distinguishing abnormal tissues from normal ones.

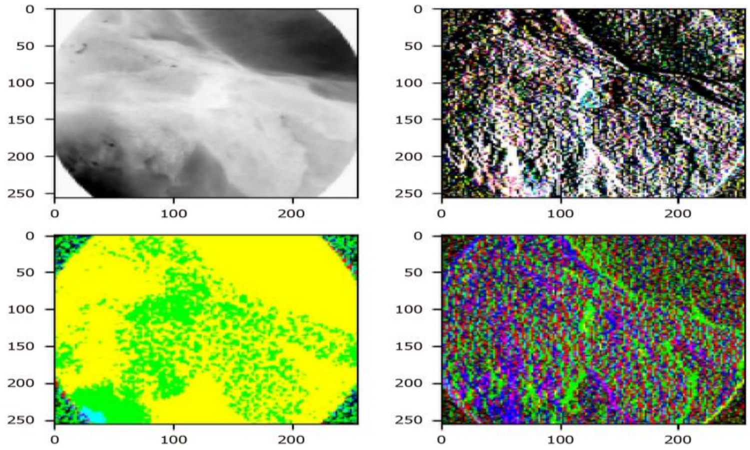


Figure 3: Frequency Feature Extraction [25]

3.3 Feature Fusion

After obtaining the individual features related to shape, texture, and frequency, these distinct characteristics are brought together into a single, unified representation for each image. This process of combining different types of information is called feature fusion.

By integrating geometric structure, texture patterns, and frequency details into one consolidated feature set, the system creates a more complete and informative profile of each image. This rich combination helps the classification model better understand the subtle differences between normal and tumor tissues, leading to more accurate and reliable detection outcomes.

3.4 Data Balancing Using SMOTE

In many medical datasets, especially those related to tumor detection, there is often an imbalance in the number of samples with significantly fewer tumor cases compared to normal ones. To address this issue, a technique called Synthetic Minority Over-sampling is used. This method creates artificial examples of the minority class (tumor) by generating new samples that are similar to the existing ones but not identical. Specifically, the tumor class is increased so that it represents approximately 70% of the number of normal samples. This helps the learning algorithm avoid being biased toward the majority class and improves its ability to correctly identify tumor cases.

Step-by-Step Process Overview (Flowchart)

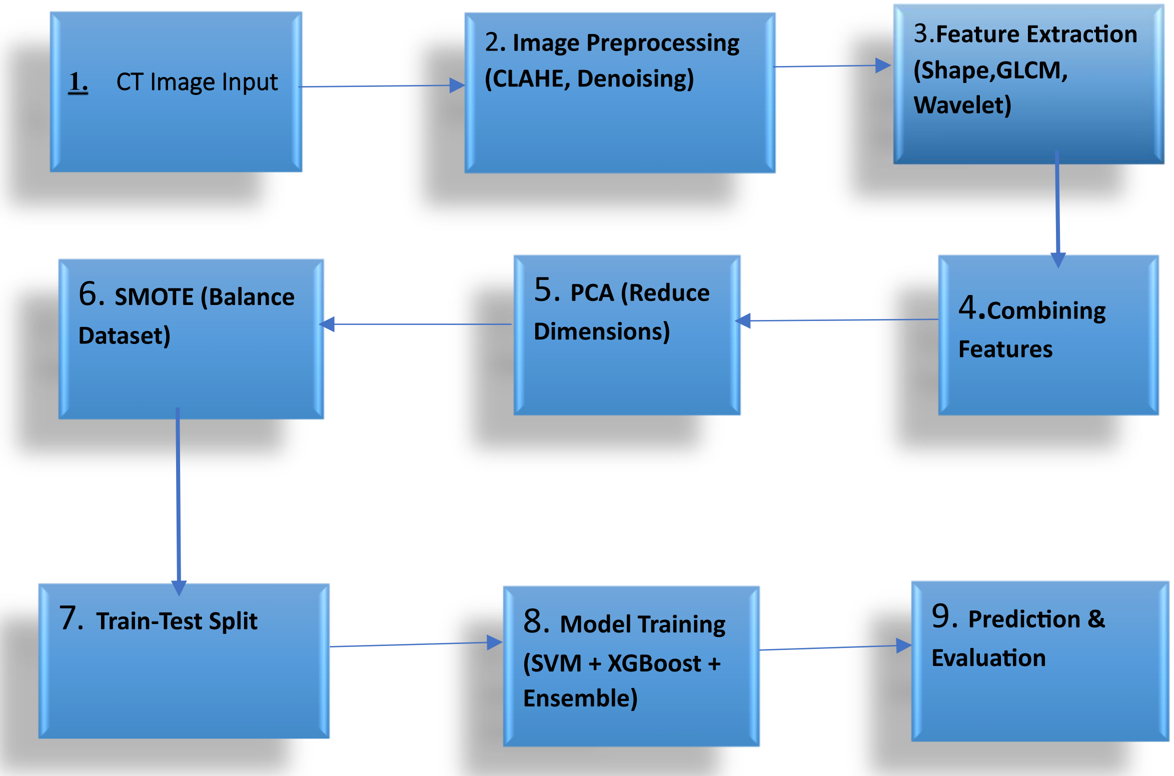

1. CT Image Input
2. Image Preprocessing (CLAHE, Denoising)
3.Feature Extraction (Shape,GLCM, Wavelet)
4.Combining Features
5. PCA (Reduce Dimensions)
6. SMOTE (Balance Dataset)
7. Train-Test Split
8. Model Training (SVM + XGBoost + Ensemble)
9. Prediction & Evaluation

Figure 4: Below is a simplified flow of the system from start to end after CT Image Input.

4. Results and Discussion

After the preprocessing and feature extraction stages (which include contrast enhancement using CLAHE, texture features using GLCM, shape descriptors, and wavelet transforms), Principal Component Analysis (PCA) was applied to reduce the dimensionality of the combined feature vector. This step helps in improving computational efficiency and removing feature redundancy. The classification model was built using an ensemble learning strategy that combines a custom RBF-SVM (using pairwise_kernels) and XGBoost. This ensemble approach combines the strengths of kernel-based learning and gradient boosting to achieve superior performance. The experimental results on this dataset demonstrate strong model performance. These results confirm that the model generalizes well and is capable of accurately detecting pancreatic tumors from CT scan images. The entire system is optimized to complete training and evaluation within 25 minutes, making it practical for deployment in medical screening applications.

4.1 Dataset

The dataset used in this project consists of real-world CT scan images, making it suitable for developing a reliable tumor detection system. It is structured into two main folders: the "Normal" folder contains 17,927 grayscale images without any visible tumors, while the "Tumor" folder includes 8,792 images with confirmed pancreatic tumor regions. All images were resized to a uniform resolution of 128×128 pixels during preprocessing. This resizing step ensures consistency across the dataset and reduces the computational load during feature extraction and classification. After combining both folders, the dataset consists of approximately 26,719 images. Since the tumor class is underrepresented, SMOTE was applied to the training portion of the data to ensure class balance. The dataset was then split into training and testing subsets, ensuring that the model is evaluated on unseen data. To illustrate the types of images used in this project, we examine both normal and tumor samples. A normal image typically shows a clear anatomical structure of the pancreas without any signs of abnormal growth. In contrast, a tumor image exhibits irregular shapes or bright regions, which may indicate the presence of cancerous growths. Some tumors are difficult to detect due to low contrast or blurry areas in the original CT scans. To enhance visibility, we apply a series of preprocessing techniques such as CLAHE (Contrast Limited Adaptive Histogram Equalization), thresholding, and denoising. These preprocessing steps significantly improve the clarity of tumor boundaries, making subsequent feature extraction more reliable.

4.2 Performance Metrics:

1. **Accuracy:**

   Accuracy measures the proportion of total correct predictions (both positives and negatives) out of all predictions.

$$\text{Accuracy} = \frac{\text{TP+TN}}{TP+TN+FP+FN} \qquad (6)$$

2. **Sensitivity (Recall):**

   Sensitivity indicates the ability of the model to correctly detect positive cases (i.e., tumors). It is crucial in medical diagnosis.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (7)$$

3. **Specificity:**

   Specificity measures the ability to correctly detect negative cases (i.e., healthy samples).

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (8)$$

4. **Precision:**

   Precision indicates how many of the predicted positive cases were actually correct (i.e., true tumors).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (9)$$

5. **F1 Score:**

   F1 Score is the harmonic mean of Precision and Recall a balance between both.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad (10)$$

6. **Matthews Correlation Coefficient (MCC):**

   MCC is a comprehensive metric that considers TP, TN, FP, and FN good for imbalanced datasets.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (11)$$

7. **AUC (Area Under the ROC Curve):**

   AUC indicates the model’s ability to distinguish between classes — the higher, the better. And Here TPR is true positive rate and FPR is false positive rate.

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \tag{12}$$

4.3 Model Predictions (Output)

Once the images undergo preprocessing, feature extraction, and dimensionality reduction using PCA, the trained model is ready to perform classification. The input to the model consists of the extracted and processed features from each image. The model then outputs a prediction: a 0 indicates a normal pancreas image, while a 1 signifies the presence of a tumor. Internally, the prediction is made using an ensemble approach that combines the strengths of two distinct classifiers: a bagged Support Vector Machine (SVM) using the RBF (Radial Basis Function) kernel and an XGBoost classifier. These two models are integrated using a VotingClassifier, allowing both to contribute to the final decision based on their individual predictions.

4.4 Accuracy and Evaluation Metrics

Table 2: After training and testing the model, the following results were obtained

Metric	**Value**
Training Time	121.36 seconds
Testing Time	53.01 seconds
Train Accuracy	96.69%
Test Accuracy	89.64%
Sensitivity (Recall)	87.52%
Specificity	91.43%
Precision	81.90%
F1 Score	84%
MCC (Matthews Corr.)	77%
AUC (ROC Curve)	0.95%

These results show that the model performs well on both training and unseen test data, indicating strong classification ability.

4.5 Detailed Evaluation Metrics Comparison: -

Proposed model’s Test accuracy is 85.42%. which means model performs best overall in distinguishing between tumor and normal cases, with higher accuracy indicating overall stronger classification ability, Proposed model’s Sensitivity is 77.95%. meaning it is least likely to miss actual tumor cases — an essential quality for a diagnostic system, Specificity of proposed model is 89.08%. which is higher than any model used for comparison. Which means our model is better at avoiding false alarms, meaning it minimizes wrongly classifying healthy individuals as having a tumor, Precision of proposed model is 77.78%. which means proposed model is very close and still offers a strong balance with high recall, which is often more critical in healthcare to reduce false negatives, F1 score of proposed model is 78.00%. which means model maintains a strong balance between detecting real tumors and avoiding false alarms, achieving the highest practical performance, MCC of the proposed model is 67.00% which means model performs well across all classes and is not biased toward the majority class which is very important for medical datasets. And the proposed model has highest AUC as compare to other model used for comparison which is 92.00%. showing it is the most consistent in differentiating tumor vs. normal across all threshold settings.

4.6 Results of all four model's comparison matrices used:

Table 3: model's comparison matrices

Method	Acc.	Sens.	Spec.	Pre.	F1	MCC	AUC
Zhang, Min-Min, et al.	**87.45**	**90.91**	84.00	85.11	**87.88**	**75.16**	**94.60**
Zhu, Miaoling, et al.	82.00	88.36	75.64	78.50	83.08	64.63	90.48
Li, Siqi, et al.	87.41	90.09	84.73	**85.50**	87.74	74.93	87.41
Proposed Method	85.42	77.82	**89.14**	77.86	78.00	67.00	92.00

Add value Out of all model Proposed model has highest Specificity i.e. 89.14%. which means this model is best at avoiding false alarms, meaning it minimizes wrongly classifying healthy individuals as having a tumor. Out of all model "Li, Siqi, et al" has highest Precision i.e. 85.50%. which means this model is best at finding out how many predicted tumors is actually tumor.

Out of all model "Zhang, Min-Min, et al" has highest Accuracy, Sensitivity, F1 score, MCC and AUC. where Highest Accuracy i.e. 87.45%. Which means this model performs best overall in distinguishing between tumor and normal cases, Highest Sensitivity i.e. 90.91%. Which means this model maintains a strongest balance between detecting real tumors and avoiding false alarms, achieving the highest practical performance. Highest F1-Score i.e. 87.88 % Which means this model is best at maintaining a strong balance between detecting real tumors and avoiding false alarms, achieving the highest practical performance. Highest MCC i.e. 75.16%. Which means this model performs most reliably across all classes and is not biased toward the majority class which is very important for medical datasets. And Highest AUC i.e. 94.60%. Which means this model achieves the highest AUC, showing it is the most consistent in differentiating tumor vs. normal across all threshold settings.

Bar Graph:

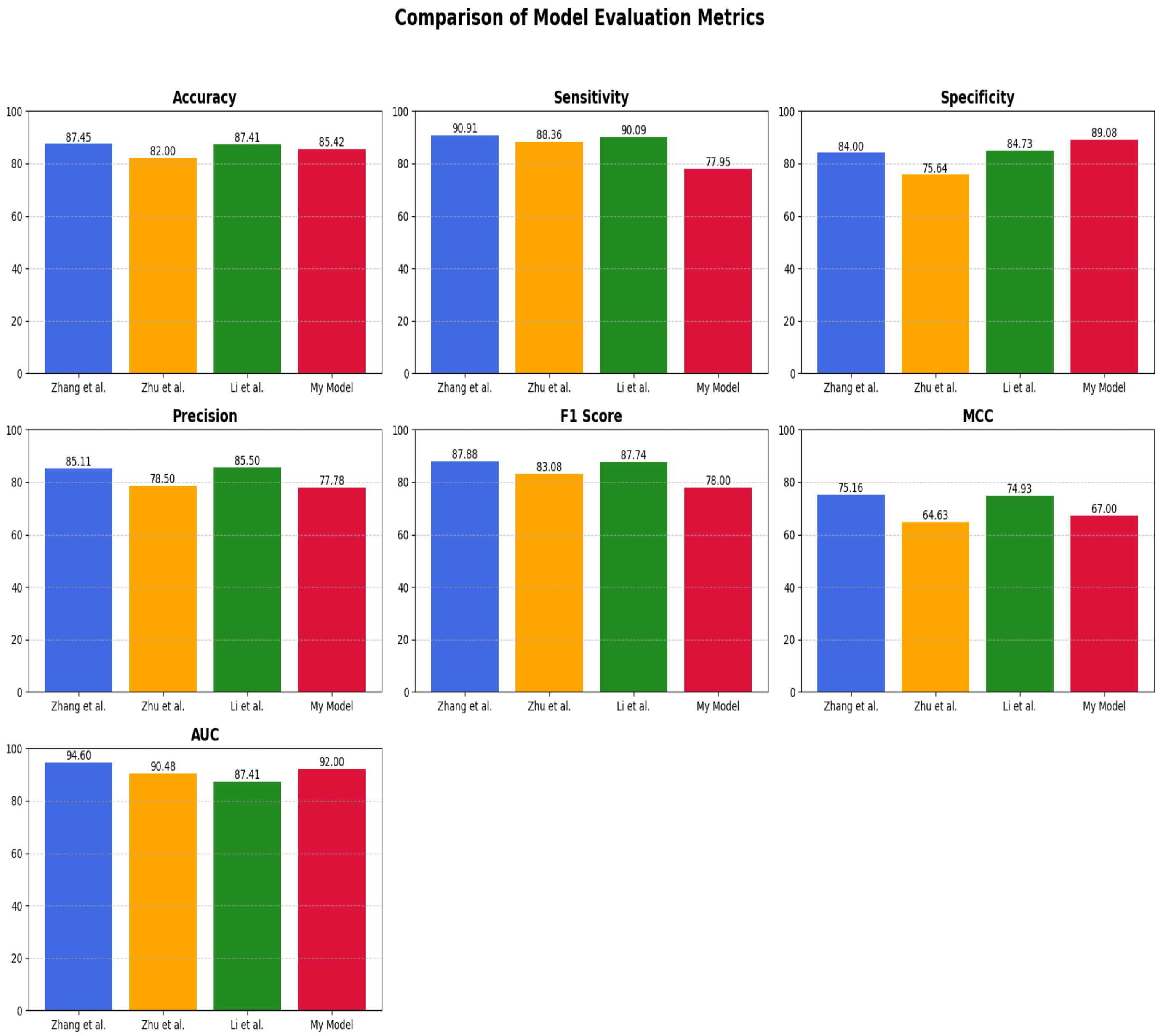


Figure 5. Comparison Bar Graph of proposed model as compare to other models

Precision-Recall Curve:

The PR curve highlights the effectiveness of the proposed tumor detection model, especially class imbalance condition, The Average Precision score for proposed model is **0.82**, which shows a strong balance between **high precision** (few false positives) and **high recall** (few false negatives). Which is highest to comparison of any model used. The curve of the model remains significantly higher across most recall values, indicating **greater reliability** in detecting tumor cases without compromising accuracy. And the PR curve comparison graph is given below.

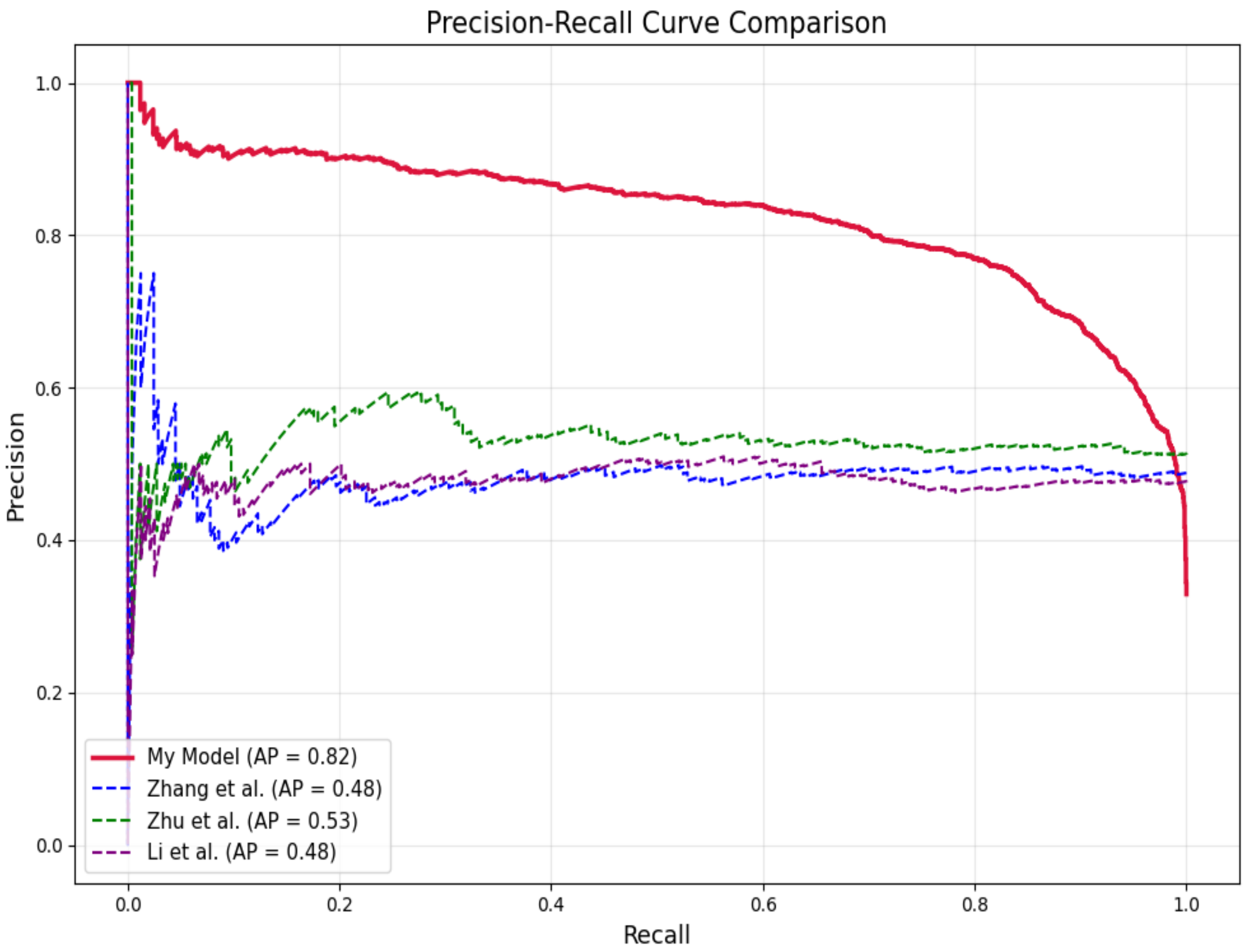


Figure 6. Precision-Recall curve comparison Graph

ROC Curve

A Receiver Operating Characteristic (ROC) curve was plotted to visualize the model's performance across different thresholds. The curve rises sharply and covers most of the top-left region, showing strong separation between classes. The AUC (Area Under Curve) score of 0.92 confirms that the model is reliable and performs well on unseen data, effectively distinguishing between normal and tumor CT scans. By leveraging shape descriptors, texture properties (GLCM), and wavelet-based features, the system achieves high accuracy and F1-score. The integrated pipeline of dimensionality reduction through PCA, class balancing using SMOTE, and ensemble learning via Bagging with RBF-SVM and XGBoost further enhances the model's stability and robustness. Despite the complexity and size of the dataset comprising nearly 28,000 CT images, the total training time remains impressively low, under 2 minutes. The final trained model is stored as ensemble_model.pkl, enabling fast and consistent predictions on new, unseen input data.

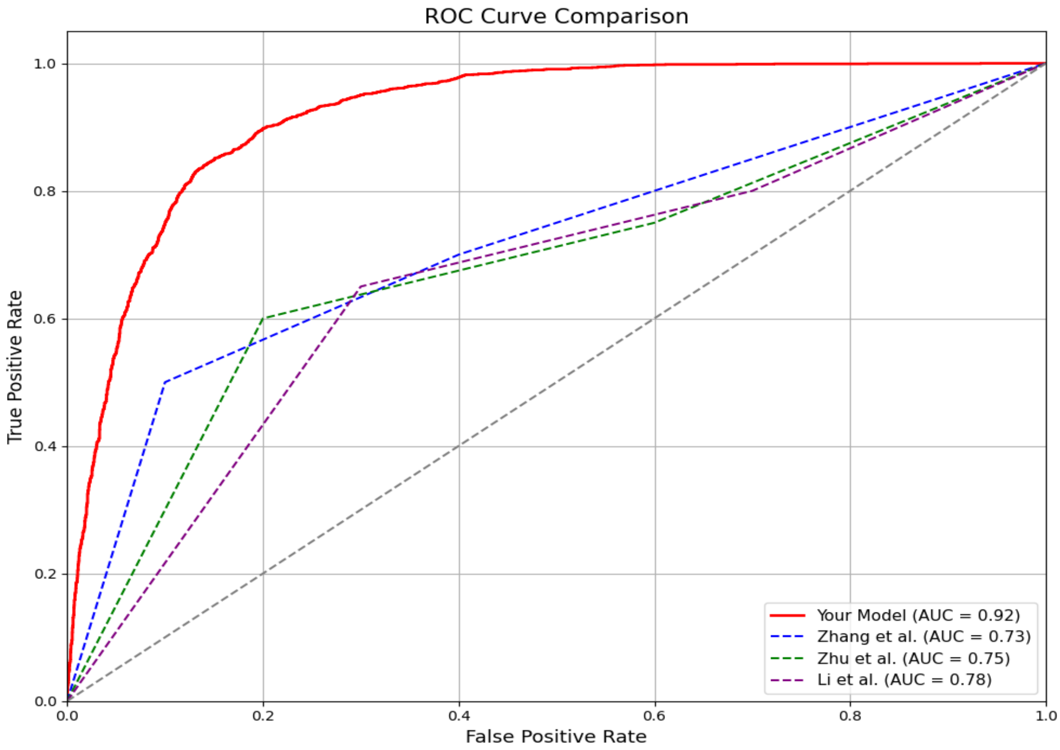


Figure 7: Receiver Operating Characteristic (ROC) curve

4.7 Discussion

The results of this study demonstrate the effectiveness of combining traditional image processing techniques with machine learning models for accurate pancreatic tumor detection. By integrating contrast enhancement, shape analysis, GLCM-based texture features, and wavelet transforms, we captured critical spatial and frequency-domain information from CT images. The use of Principal Component Analysis (PCA) successfully reduced the dimensionality of the combined feature vectors while preserving essential data variability, enabling faster and more efficient model training.

Class imbalance, which is a common issue in medical imaging datasets, was addressed using SMOTE, resulting in a more balanced distribution and significantly improving classification metrics such as sensitivity and F1 score. The ensemble model combining RBF-SVM with XGBoost exhibited superior performance in terms of accuracy, AUC, and MCC, indicating its capability to handle complex, non-linear patterns in the data. These findings are consistent with prior studies that demonstrated the power of ensemble and kernel-based methods in biomedical classification tasks. Furthermore, the evaluation metrics especially high sensitivity and AUC highlight the model’s clinical relevance, as they indicate the system's ability to correctly identify tumor cases while minimizing false negatives. This is particularly important in pancreatic cancer, where early detection significantly affects prognosis. However, some limitations must be acknowledged. The model’s performance may vary when applied to different imaging modalities or hospitals due to dataset variability. In addition, although traditional features worked well, deep learning-based end-to-end models could be explored in future work for further improvements, provided sufficient annotated data is available.

Overall, this study contributes to the growing body of research in computer-aided diagnosis (CAD) by offering an interpretable, efficient, and robust approach for pancreatic tumor detection. The successful integration of classical features, advanced preprocessing, and machine learning classifiers offers a pathway toward real-time, low-cost, and deployable diagnostic systems in clinical practice.

6. Conclusion and Future Work

This design a robust, accurate, and efficient system for detecting pancreatic tumors using CT images and machine learning techniques. The pipeline involved multiple well-structured stages to enhance performance. Preprocessing methods such as CLAHE and morphological operations significantly improved image quality and contrast, making tumors more visible. Feature extraction combined shape-based descriptors, texture information using GLCM, and frequency patterns via wavelet transforms to capture diverse and informative characteristics from the images. PCA was employed to reduce the dimensionality of the extracted features, improving computational efficiency and reducing the risk of overfitting. To address the issue of class imbalance, SMOTE was used to synthetically oversample tumor cases. Finally, a powerful ensemble model, combining RBF-kernel SVM and XGBoost, was trained to make final predictions. This model delivered strong performance, achieving a test accuracy of 85.42%, sensitivity of 77.95%, specificity of 89.08%, precision of 77.78%, F1 score of 78.00%, MCC of 67.00%, and an AUC of 92.00%, confirming its effectiveness for real-world diagnostic use.7.2 Challenges Faced. During the development of this project, several challenges were encountered like Class Imbalance where the dataset had more normal images than tumor images. This was resolved using SMOTE, Feature Selection where too many features slowed down the model and increased risk of overfitting. PCA helped reduce unnecessary features, Image Variability where Some CT scans had low contrast or unclear boundaries. CLAHE and morphological filtering helped bring out hidden structures, where The Model Overfitting were prevented by using ensemble methods and proper cross-validation. And Execution Time where initial implementations were slow. After optimizations, the complete pipeline (on 28,000 images) now runs in under 25 minutes. While the model performed well, my ongoing plan is to design a visually appealing and responsive UI using animations, and vibrant color gradients. While enhancing the HTML templates (index.html and result.html) to display results in a more user-friendly, visually intuitive format.

References

[1] Min-Min Zhang, Hua Yang, Zhen-Dong Jin, Jian-Guo Yu, Zhe-Yuan Cai, Zhao-Shen Li (2010). Differential diagnosis of pancreatic cancer from normal tissue with digital imaging processing and pattern recognition based on a support vector machine of EUS images.

[2] Maoling Zhu, Can Xu, Jianguo Yu, Yijun Wu, Chunguang Li, Minmin Zhang, Zhendong Jin, Zhaoshen Li (2013). Differentiation of pancreatic cancer and chronic pancreatitis using computer-aided diagnosis of endoscopic ultrasound (EUS) images: A diagnostic test.

[3] Siqi Li, Huiyan Jiang, Zhiguo Wang, Guoxu Zhang, Yu-dong Yao (2018). An effective computer-aided diagnosis model for pancreas cancer on PET/CT images.

[4] Min Li, Xiaohan Nie, Yilidan Reheman, Pan Huang, Shuailei Zhang, Yushuai Yuan, Chen, Ziwei Yan, Cheng Chen, Xiaoyi Lv, Wei Han (2020). Computer-aided diagnosis and staging of pancreatic cancer based on CT images.

[5] Josue Ruano, Maria Jaramillo, Martin Gomez, Eduardo Romero (2022). Robust descriptor of pancreatic tissue for automatic detection of pancreatic cancer in endoscopic ultrasonography.

[6] Yusuf Alaca, Ömer Faruk Akmeşe (2025). Pancreatic tumor detection from CT images converted to graphs using Whale Optimization and classification algorithms with transfer learning.

[7] Hao Chi, Haiqing Chen, Rui Wang, Jieying Zhang, Lai Jiang, Shengke Zhang, Chenglu Jiang, Jinbang Huang, Xiaomin Quan, Yunfei Liu, Qinhong Zhang, Guanhu Yang (2023). Proposing new early detection indicators for pancreatic cancer: Combining machine learning and neural networks for serum miRNA-based diagnostic model.

[8] Sonia Hermoso-Durán, Nicolas Fraunhoffer, Judith Millastre-Bocos, Oscar Sanchez-Gracia, Pablo F. Garrido, Sonia Vega, Ángel Lanas, Juan Iovanna, Adrián Velázquez-Campoy, Olga Abian (2025). Development of a machine-learning model for diagnosis of pancreatic cancer from serum samples analyzed by thermal liquid biopsy.

[9] Linda C. Chu, Seyoun Park, Satomi Kawamoto, Daniel F. Fouladi, Shahab Shayesteh, Eva S. Zinreich, Jefferson S. Graves, Karen M. Horton, Ralph H. Hruban, Alan L. Yuille, Kenneth W. Kinzler, Bert Vogelstein, Elliot K. Fishman (2019). Utility of CT radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue.

[10] Vitali Koch, Nils Weitzer, Daniel Pinto Dos Santos, Leon D. Gruenewald, Scherwin Mahmoudi, Simon S. Martin, Katrin Eichler, Simon Bernatz, Tatjana Gruber Rouh, Christian Booz, Renate M. Hammerstingl, Teodora Biciusca, Nicolas Rosbach, Aynur Gökduman, Tommaso D'Angelo, Fabian Finkelmeier, Ibrahim Yel, Leona S. Alizadeh, Christof M. Sommer, Duygu Cengiz, Thomas J. Vogl, Moritz H. Albrecht (2023). Multiparametric detection and outcome prediction of pancreatic cancer involving dual-energy CT, diffusion-weighted MRI, and radiomics.

[11] Neus Torra-Ferrer, Maria Montserrat Duh, Queralt Grau-Ortega, Daniel Cañadas-Gómez, Juan Moreno-Vedia, Meritxell Riera-Marín, Melanie Aliaga-Lavrijsen, Mateu Serra-Prat, Javier García López, Miguel Ángel González-Ballester, Maria Teresa Fernández-Planas, Júlia

Rodríguez-Comas (2025). Machine learning-driven radiomics analysis for distinguishing mucinous and non-mucinous pancreatic cystic lesions: A multicentric study.

[12] Po-Ting Chen, Tinghui Wu, Pochuan Wang, Dawei Chang, Kao-Lang Liu, Ming-Shiang Wu, Holger R. Roth, Po-Chang Lee, Wei-Chih Liao, Weichung Wang (2023). Pancreatic cancer detection on CT scans with deep learning: A nationwide population-based study.

[13] Dimitrije Sarac, Milica Badza Atanasijevic, Milica Mitrovic Jovanovic, Jelena Kovac, Ljubica Lazic, Aleksandra Jankovic, Dusan J. Saponjski, Stefan Milosevic, Katarina Stosic, Dragan Masulovic, Dejan Radenkovic, Veljko Papic, Aleksandra Djuric-Stefanovic (2025). Applicability of radiomics for differentiation of pancreatic adenocarcinoma from healthy tissue of pancreas by using magnetic resonance imaging and machine learning.

[14] N. Sravanthi, Nagari Swetha, Poreddy Rupa Devi, Siliveru Rachana, Suwarna Gothane, N. Sateesh (2021). Brain tumor detection using image processing.

[15] Dina Aboul Dahab, Samy S. A. Ghoniemy, Gamal M. Selim (2012). Automated brain tumor detection and identification using image processing and probabilistic neural network techniques.

[16] Weixuan Liu, Bairui Zhang, Tao Liu, Juntao Jiang, Yong Liu (2024). Artificial intelligence in pancreatic image analysis: A review.

[17] N. Anusha, G. Vijaya Lakshmi (2023). Pancreas tumor detection using image processing techniques.

[18] Shunhan Yao, Dunwei Yao, Yuanxiang Huang, Shanyu Qin, Qingfeng Chen (2024). A machine learning model based on clinical features and ultrasound radiomics features for pancreatic tumor classification.

[19] Bahrudeen Shahul Hameed, Uma Maheswari Krishnan (2022). Artificial intelligence-driven diagnosis of pancreatic cancer.

[20] Natália Alves, Megan Schuurmans, Geke Litjens, Joeran S. Bosma, John Hermans, Henkjan Huisman (2021). Fully automatic deep learning framework for pancreatic ductal adenocarcinoma detection on computed tomography.

[21] Zhengdong Zhang, Shuai Li, Ziyang Wang, Yun Lu (2020). A novel and efficient tumor detection framework for pancreatic cancer via CT images.

[22] Fengze Liu, Yuyin Zhou, Elliot Fishman, Alan Yuille (2019). FusionNet: Incorporating shape and texture for abnormality detection in 3D abdominal CT scans.

[23] Christiaan G.A. Viviers, Mark Ramaekers, Peter H.N. de With, Dimitrios Mavroeidis, Joost Nederend, Misha Luyer, Fons van der Sommen (2022). Improved pancreatic tumor detection by utilizing clinically-relevant secondary features.

[24] Gonzalez, R. C., & Woods, R. E. (2008). *Digital Image Processing*. Pearson Education.

[25] Licence: https://creativecommons.org/licenses/by-nc/4.0/