UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

# PREDICTING KIDNEY POST-TRANSPLANTATION FUNCTION FROM OPTICAL COHERENCE TOMOGRAPHY IMAGES USING MACHINE LEARNING APPROACHES

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
MASTER OF SCIENCE

By

BORNFACE MUTEMBEI
Norman, Oklahoma
2025

PREDICTING KIDNEY POST-TRANSPLANTATION FUNCTION FROM OPTICAL COHERENCE TOMOGRAPHY IMAGES USING MACHINE LEARNING APPROACHES

A THESIS APPROVED FOR THE
STEPHENSON SCHOOL OF BIOMEDICAL ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Qinggong Tang (Chair)

Dr. David Miller

Dr. Handan Acar

## Acknowledgments

First, I would like to express my sincere gratitude to my advisor, Dr. Qinggong Tang, for their invaluable support throughout my time in the Biophotonics Laboratory and for their mentorship during this project. Dr. Tang cultivated an environment that encouraged both independent work and collaboration.

I am also thankful to my committee members, Dr. Handan Acar and Dr. David Miller for their support during the revision of this work. In addition to serving on my committee, Dr. Acar has provided thoughtful guidance as a faculty mentor in matters of career development and professional growth.

A special thanks is extended to Dr. Theodore Trafalis from the School of Industrial Systems Engineering at the University of Oklahoma for insightful advice and mentorship throughout the course of this work.

Lastly, I am grateful to my grandparents Samson and Harriet Muthamia whose unwavering support and investment in my education has made this journey possible.

# Table of Contents

# List Of Tables

# List Of Figures

# Abstract

Delayed graft function (DGF) is a major post-transplant complication in kidney recipients, particularly from deceased donors. This study investigates the effectiveness of machine learning approaches in predicting DGF using texture features extracted from optical coherence tomography (OCT) images, supplemented with Kidney Donor Profile Index (KDPI) scores. To address significant class imbalance between immediate graft function (IGF) and DGF, three strategies were evaluated: threshold optimization using the GHOST algorithm, data balancing via SMOTE-Tomek, and cost-sensitive learning (CSL). We demonstrate that classifiers trained on KDPI scores alone underperformed compared to those trained OCT images derived texture features. Classifiers trained on KDPI alone showed moderate improvements with GHOST, but incorporating OCT-derived texture features significantly enhanced model performance across all classifiers. These findings underscore the utility of OCT imaging in assessing kidney allograft quality and predicting post-transplant outcomes, highlighting the potential of machine learning classifiers to estimate the risk of delayed graft function (DGF) in deceased donor kidneys prior to transplantation.

# Chapter 1: Introduction

## 1.1 Donor Categories and Expanded Criteria Donors

Kidney transplants afford patients with end-stage renal disease (ESRD) long-term, dialysis free and cost savings compared to maintenance dialysis [1]. In 2020, 130,522 people in the United States started the treatment for ESRD, with 808,000 people in total living with ESRD, 69% of were on dialysis and the rest were living with a kidney transplant [2]. To address the changing demographics of kidney donors to older populations, longer wait times and shortage of organ supply for transplantation, the Extended Criteria Donor (ECD) was introduced in March 2001[3, 4, 5]. ECD kidneys are those donated by deceased donors $\geq 60$ or $\geq 50$ years old who meet at least two of the following criteria: cerebrovascular cause of death, terminal serum creatinine levels and a history of hypertension [6].

Compared to kidneys from standard criteria donors (SCD), kidneys from expanded criteria donors (ECD) exhibit poorer long-term outcomes, rendering them unsuitable for younger ERSD patients [6]. However, they may be suitable for older patients, who experience better survival rates with ECD kidneys than with remaining on a waitlist for SCD kidneys [6].

## 1.2 Post-transplant categories, IGF versus DGF

A common early post-transplant complication and cause kidney allograft failure is delayed graft function (DGF) defined as the need for dialysis one week post transplantation and it is increased in ECD versus SCD kidney transplants [7]. In the short-term, DGF impairs the patient quality of life, necessitates dialysis and further biopsy evaluations that prolong patient's stay at the hospital while in the long-term, DGF is associated with poorer graft function 1 year post transplantation for ECD donors [8]. Potential risk factors for DGF include the mechanism of ischemia (adequate organ perfusion is required to avoid hypoxemia), donation after cardiovascular or brain death, ECD kidneys and composite donor or recipient risks[7, 8, 9].

Organs from marginal donors (including ECD) are characterized by an increased rate of DGF and primary graft nonfunction (PNF), both of which increase the risk of acute rejection and reduce post-transplant survival [10]. Organ preservation is an essential part of the transplantation process and it serves to maintain the function of the organ to ensure that the graft will function upon transplantation [11]. The common modes of preservation for kidneys are either static or dynamic, with simple cold storage (SCS) as the main method of the former and hypothermic machine perfusion (HMP) as the main method of the latter [10, 11].

In SCS, the kidney is flushed with a sterile preservation solution immediately after procurement and stored in a disposable bag covered with ice to maintain a temperature between 4° C-10°C at which cellular degradation is significantly lowered [10, 11, 12]. Hypothermic machine perfusion (HMP) involves the continuous or intermittent perfusion of the kidney with a cold preservation solution via a machine perfusion system, maintaining temperatures between 0°C and 4°C, to sustain nutrient supply and eliminate metabolic waste, ultimately extending both preservation time and quality of the kidneys [10, 11, 13]. Although SCS is a simple methods that can be applied in any hospital without extensive training or expensive equipment, its nearly impossible to ensure consistency in the parameters used to evaluate the quality of the organ and to measure ischemic damage [10].

In fact, ECD kidneys remain underutilized because SCS is unsuitable for preservation of marginal kidneys due to increased incidence of PNF. Recent studies have reported that HMP is a better preservation method, especially high-risk donors because besides prolonged preservation times, it provides allows the evaluation of organ viability and quality prior to transplantation [13, 14, 15].

## 1.3 Utilizing OCT to assess renal graft function

Optical coherence tomography (OCT) is a non-invasive optical imaging modality with a millimeter depth resolution for real-time, non-invasive, depth-resolved imaging of biological tissues in situ [16, 17, 18]. With typical image resolutions of 1-15µm [18], OCT has been used to evaluate kidney microstructure [18, 19, 20, 21, 22] and to predict post-transplant renal function [23, 24]. OCT has the potential to provide comprehensive imaging of multiple regions across the kidney surface, enabling a thorough evaluation of renal grafts [23]. This offers a non-invasive alternative to the current gold standard biopsy evaluation which can cause tissue damage that mimics injury from ischemia and other pathological conditions [25]. While some studies have reported that histological scores from donor kidney biopsies correlate with outcomes such as delayed graft function (DGF) or initial graft function (IGF), the systematic review by [26] found these associations to be highly inconsistent, with many studies showing no significant correlation, highlighting the conflicting predictive value of biopsy findings for post-transplant outcomes. In this study, OCT was used to obtain image data of kidney organs pre- and post-transplantation, and the resulting images were used to obtain texture parameters used to train machine learning models to predict post-transplantation function.

## 1.4 Texture Feature Extraction

In biomedical imaging, texture is defined as the spatial variation or organization of pixel values, capturing patterns such as smoothness or roughness that are not conveyed by intensity values alone [27, 28, 29] . The most popular texture extraction method involves extracting radiomics features from images acquired using various radiography modalities including computed tomography, magnetic resonance imaging, positron emission tomography, and ultrasonography [30].

First proposed in 2012, radiomic features have gained popularity in cancer detection and classification due to their ability to characterize both temporal and spatial heterogeneity in biomedical tissues, particularly in solid tumors [31, 32, 33]. Radiomics was developed to support high-throughput, quantitative analysis of the growing volume of biomedical imaging data. Humeau 2019 provides a comprehensive review of texture extraction algorithms, categorizing them into statistical, structural, transform-based, model-based, graph-based, learning-based and entropy-based approaches and recent advances in this area [34]. In previous work, we demonstrated the utility of an expanded feature set generated using the PyFeats algorithm to classify multicellular cancer spheroids (MCTS) of varying cell compositions and sizes treated with different anticancer drugs, using machine learning models [16]. The goal was to create a large dataset of texture features for comprehensive characterization of treatment-induced changes in MCTS, with potential applications beyond oncology, such as evaluating spatial heterogeneity in the quality of donor grafts across different sites on the organ in ongoing transplantation projects. However, due to a significant class imbalance between patients with delayed graft function and those with immediate graft function in our current dataset, we adopted a smaller radiomic feature set to facilitate more meaningful comparison with another ongoing project using similar data.

## 1.5 The imbalanced data problem

Imbalanced data refers to an uneven distribution of samples across different classes. In binary classification, the “minority” or “negative” class denotes the class with fewer samples, while the “majority” or “positive” class represents the one with more samples [35, 36]. In the data used in this study, the immediate graft function (IGF) category has disproportionately more samples compared to the delayed graft function (DGF). The phenomenon of class imbalance in kidney transplant outcomes, IGF versus DGF, is evident from their uneven prevalence, with IGF occurring in approximately 66.8% of kidney transplant recipients, compared to 33.2% for DGF [9, 37].Conventional classifiers often exhibit lower performance for the minority class because they prioritize overall accuracy hence favoring the majority class [36].

Additionally, conventional classifiers are designed under the assumption that the dataset is balanced and that misclassification costs are equal across all classes [38, 39, 40]. Consequently, such models tend to regard the minority class as noise and become biased towards the majority class and have lower generalization to new and unseen data [41, 42, 43]. To predict the risk of DGF, correct classification of the minority class has greater value than IGF due to its transplantation outcomes including the risk of graft failure [5]. Solutions to address class imbalance are broadly categorized into three groups: data-level, algorithm level and hybrid approaches. Data-level involves resampling (under sampling, oversampling or combined under and oversampling) to balance the number of samples from the majority and minority classes [36]. Algorithm level solutions aim to establish robust algorithms that can effectively classify the minority class mainly through cost-sensitive learning [44], ensemble learning [48] or adjusting the prediction threshold to one that optimizes the classification of imbalanced data [45].

### 1.5.1 Data-level solutions to imbalanced data

Resampling methods are commonly used to address class imbalance by adjusting the distributions of the minority and majority class by oversampling the minority class or under-sampling the majority class [40, 42]. Under-sampling reduces the size of the majority class by removing samples to match the minority class, while oversampling increases the size of the minority class by adding duplicates of its existing samples [40]. Pure under-sampling is suitable for large datasets with a lower class imbalance ratio and reduces computation time but leads to information loss from the majority class [46, 47]. On the other hand, oversampling increases computation time and may cause overfitting [40, 47]. To overcome the limitations of pure oversampling and under-sampling, modified algorithms like the Synthetic Minority Oversampling Technique (SMOTE) and the Adaptive Synthetic (ADASYN) algorithm have been developed. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class samples and their nearest neighbors in feature space, creating broader decision regions that improve model generalization while addressing class imbalance [48]. Advancements and applications of the SMOTE algorithm have been reviewed in [49], with two variations, SMOTE-Tomek and SMOTEfied-GANs, being particularly relevant to this study [43, 50, 51]. Tomek Links is a technique that identifies pairs of instances from different classes that are close to each other and removes those that contribute to class overlap [52]. SMOTE-Tomek combines SMOTE with Tomek Links by first generating synthetic samples using SMOTE and then applying Tomek Links to remove instances located near class boundaries, thus reducing overlap and improving class separation [50]. SMOTified-GAN is a hybrid approach that combines the Synthetic Minority Over-Sampling Technique (SMOTE) and Generative Adversarial Networks (GANs) to address class imbalance, where SMOTE generates synthetic minority samples that are then refined by GAN to produce more realistic data, improving sample quality and classification performance [51]. SMOTE algorithms with GAN have been utilized to address class imbalance in weather dataset by [43] outperforming the standard SMOTE implementation. In this work, we investigate whether balancing the OCT texture data using SMOTE-Tomek improves the classification accuracy of various classifiers to predict post-transplant function.

### 1.5.2 Algorithm-level Solutions to imbalanced data

Various predictive models have been trained to predict DGF including logistic regression [53]; random forest classifiers (RF), k-nearest neighbor (kNN), linear discriminant analysis, quadratic discriminant analysis, decision trees (DT), and gradient boosting (GB) [54]; and gaussian naive Bayes, multiple-layer perceptron (MLP), and eXtreme Gradient Boosting (XGB) [55]. These models vary in their susceptibility to class imbalance, with some achieving high accuracy but lower stability, while others are less affected by imbalance but exhibit reduced predictive performance [38]. In this work, we compare the performance of DT, GB, random forests (RF), Complement Naive Bayes (CNB), support vector machines (SVMs), Stochastic Gradient Descent (SGD), and MLP classifiers on both texture and kidney transplant recipient KDPI score to predict post-transplant function.

# Chapter 2: Methods

## 2.1 Patient Demographic and Transplant Groups

The dataset used in this paper was obtained from a study approved by the Institutional Review Boards of Georgetown University and the University of Maryland (study number: IRB#2010-396). Eligible included patients $\geq$ 18 years old who had received a kidney transplant at the MedStar Georgetown Transplant Hospital and all patients provided written consent prior to enrollment. Patients included in the study consisted of approximately 60% male and 40 female recipients with a median age of $52\pm12.5$ years, and a mean BMI of $28.4\pm$ 4.7. The total number of kidneys imaged in the original study was 169, with 66 from living donor kidney transplants (LDKT) versus 103 from deceased donor kidney transplants (DDKT). Only data from deceased donors were used for this study. All LDKTs were preserved by static cold storage (SCS). 88 of the DDKTs were stored by SCS, while the rest of the DDKTs were preserved by hypothermic machine perfusion (HMP). Regarding donor classification, 26 of the 88 SCS kidneys were classified as expanded donor criteria (ECD) kidneys, while the remaining 62 SCS kidneys were classified as standard criteria donor (SCD) kidneys. Only 2 of the kidneys in the HMP group were classified as ECD kidneys. In the original study, four patients were excluded from including two patients involved in anti-delayed graft function during the duration of the study and two other patients whose OCT image quality was compromised. Additional information on patient demographics and the data set is currently available in a previous article [24].

## 2.2 Data Processing

The data used to train machine learning models comprised of the recipient data (physiological data from monitoring pre-and post-transplantation) and texture data extracted from OCT images. The steps used to pre-process this data, balance the dataset, select and extract features for training the ML models are described in this section and summarized in Figure 2.1.

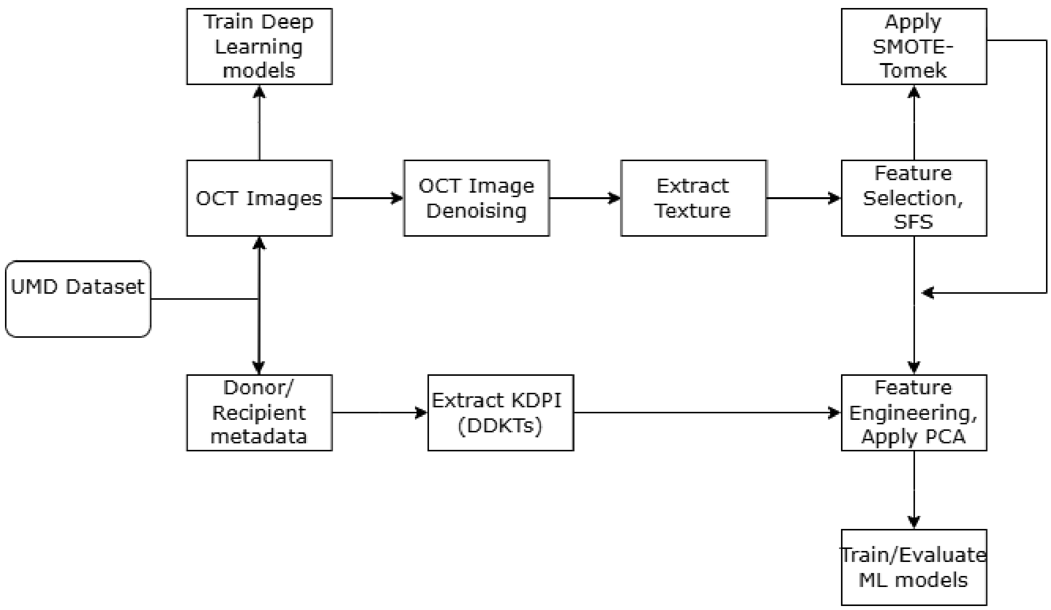



Figure 2.1: Data processing workflow outlining steps taken to prepare the data to train machine learning models.

### 2.2.1 OCT Image Acquisition

A spectral-domain OCT imaging system (Telestro-II, Thorlabs Inc.) with a 1325 nm center wavelength was used to image donor kidneys. This system had a 3.5 mW incident power, 36 mm focal length (LSMO3, Thorlabs Inc) with a lateral and axial resolution of 13 µm and 5.5 um. The system's field of view in x and z axis was 4.9

and 1.9 mm. Additional specification of the imaging system used can be found in our previous work [24].

### 2.2.2 OCT Image Segmentation

I trained four deep learning models (Resnet50, U-net, Pixel to Pixel, and the Segment Anything model) to segment the background from the area of interest (ROI). This step ensures that the texture data extracted is from the tissues of the kidney and not the background. I formulated this task as an image-to-image translation task where a model learns how to map input images into output images, for example, OCT images with background noise to binary masks. Predicting binary masks ensured signal integrity in the region of interest after segmenting the OCT images. First, 4370 OCT images were randomly selected (training = 3710 and validation = 660) and manually labeled on ImageJ. By applying triangle thresholding and flood filling the manually labeled images, I obtained the binary masks. These binary masks denote the background and ROI were used as the target output when training the deep learning models.

The Pixel to pixel (Pix2pix) is a conditional generative adversarial network (cGAN) proposed by Isola et al [56] suitable for image translation tasks with various applications: generating maps from satellite images [57], image colorization [58] and binary image denoising [59]. GANs are suitable for image-to-image translation tasks because they allow generalizability across new image domains instead of training task-specific models. The Pix2pix model architecture comprises a U-net-based image generator [60] and a convolutional PatchGAN discriminator. The U-net model consists of an upsampler and a downsampler with unique skip connections that allow high segmentation accuracy with fewer training images [60]. The generator and discriminator are pitted against each other with the generator predicting fake image patches that resemble the target output and the discriminator classifying those patches as real or fake. Training both components simultaneously ensures that the generator learns how to predict output images that closely resemble the target images that are classified by the discriminator as real.

I compared the performance of Pix2pix to that of the decoupled U-net generator, Resnet50, and Segment Anthing Model (version 1, [61]). The pix2pix, Resnet50, and U-net models were trained on Python (Python Org, 3.12) but the SAM model was pre-trained and available for validation via the ImageSegmenter App on MATLAB (MATLAB, R2024b. Natick, Massachusetts). All models were validated using the unseen 660 validation images to compute the metrics tabulated in Table 2.1.

The accuracy metric assesses binary pixel accuracy while the intersection over the union (IOU) was used to assess how accurately the models segmented the ROI. On the other hand, the mean square error (MSE) was used to quantify how the model misclassifies pixels in the background and ROI in the predicted images versus the ground truth. The structural similarity index metric (SSIM) was used to quantitatively assess image quality in luminance, contrast, and structure. Thus, SSIM overcomes the limitations of MSE which assumes statistical features in the image are spatially stationary [62]. Lastly, the peak-to-signal ratio (PSNR) was used to assess the quality of image segmentation with a higher PSNR score indicating higher quality of predicted binary masks. The Pix2pix model outperformed the decoupled U-net model generator, Resnet50, and Segment Anything Model (SAM) as shown in Figure 2.2 and Table 2.1.

Table 2.1: Comparison of deep learning models for denoising OCT images

| Models | Acc. | IOU | SSIM | PSNR | MSE |
|---|---|---|---|---|---|
| Resnet50 | 0.9546 | 0.8915 | 0.9235 | 14.0000 | 2979.1380 |
| U-Net | 0.9796 | 0.9497 | 0.9506 | 17.4769 | 1567.4580 |
| pix2pix | 0.9780 | 0.9457 | 0.9525 | 17.1779 | 1542.5190 |
| SAM | 0.9572 | 0.9032 | 0.9415 | 15.2344 | 2798.1550 |

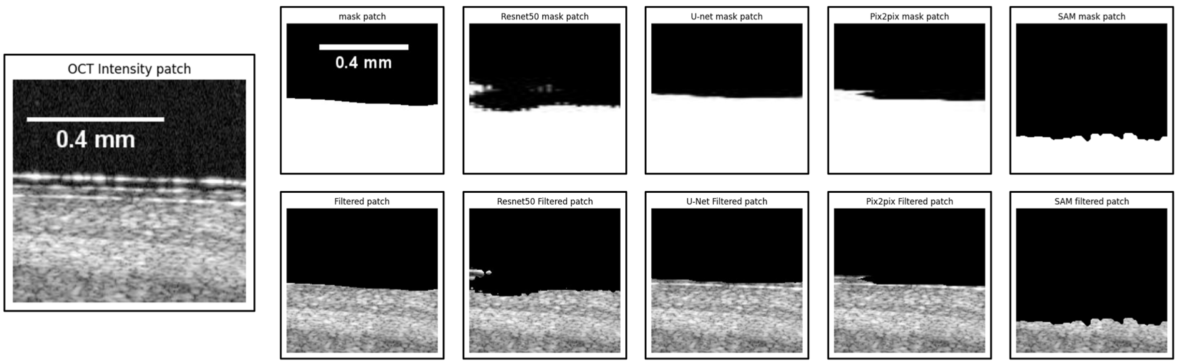



Figure 2.2: Comparison of mask prediction and image filtering for PS-OCT intensity images by Sam, Resnet50, U-net, and Pix2Pix models. SAM, Segment Anything Model.

### 2.2.3 Texture Extraction

Radiomic features encompass measurements that describe the distribution of intensity values, spatial arrangements of pixels, patterns of texture heterogeneity, as well as the shape, morphology, and volume of the imaged structures. Figure 2.3 presents a hierarchically clustered heatmap of Pearson correlation coefficients for the 1,373 extracted radiomic features, with values ranging from -1 to 1 [63, 64]. Negative correlations are depicted in brown, and positive correlations in green, with darker shades indicating stronger correlations. Overall, the heatmap reveals generally low inter-feature correlations, and the accompanying dendrogram suggests that most associations occur between features from different categories.

Table 2.2: Radiomics texture feature extraction from various image types

| Image Type | Shape | First order | GLCM | GLDM | GLRLM | GLSZM | NGTDM | Total |
|---|---|---|---|---|---|---|---|---|
| Original | 14 | 18 | 24 | 14 | 16 | 16 | 5 | 107 |
| Exponential | 0 | 16 | 0 | 7 | 8 | 3 | 0 | 34 |
| Gradient | 0 | 18 | 24 | 14 | 16 | 16 | 5 | 93 |
| LBP3D | 0 | 45 | 21 | 14 | 16 | 16 | 5 | 117 |
| Logarithm | 0 | 18 | 24 | 14 | 16 | 16 | 5 | 93 |
| Square | 0 | 17 | 24 | 14 | 16 | 16 | 5 | 92 |
| Square root | 0 | 18 | 24 | 14 | 16 | 16 | 5 | 93 |
| Wavelet | 0 | 144 | 192 | 112 | 128 | 128 | 40 | 744 |
| Total | 14 | 294 | 333 | 203 | 232 | 227 | 70 | 1373 |

### 2.2.4 Feature Selection and dimension reduction

Missing values imputation was done for both the recipient and texture data prior to feature selection. In the recipient, data, we first identified the features with missing data above a threshold of 15% removed them from the dataset. The missing values in the remaining features were replaced by the mean value of all the patients. However, for the texture data, missing values were replaced with zeros. Feature selection is the process of identifying relevant features from a collection of features based on a criterion. Feature selection has the potential to reduce the dimensionality of the data, computation time and overcome overfitting resulting in improved performance of predictive models [65]. Feature selection methods include filters, wrappers, and embedded approaches. Filters evaluate features independently of the classifier, while embedded methods integrate feature selection into the learning algorithm [66]. Wrappers such as sequential feature selection (SFS), assess feature subsets using a specific learning model to minimize classification errors [66].

In this study, I applied sequential feature selection in the forward direction using random forest regressor as the core classifier to identify optimal feature subsets.

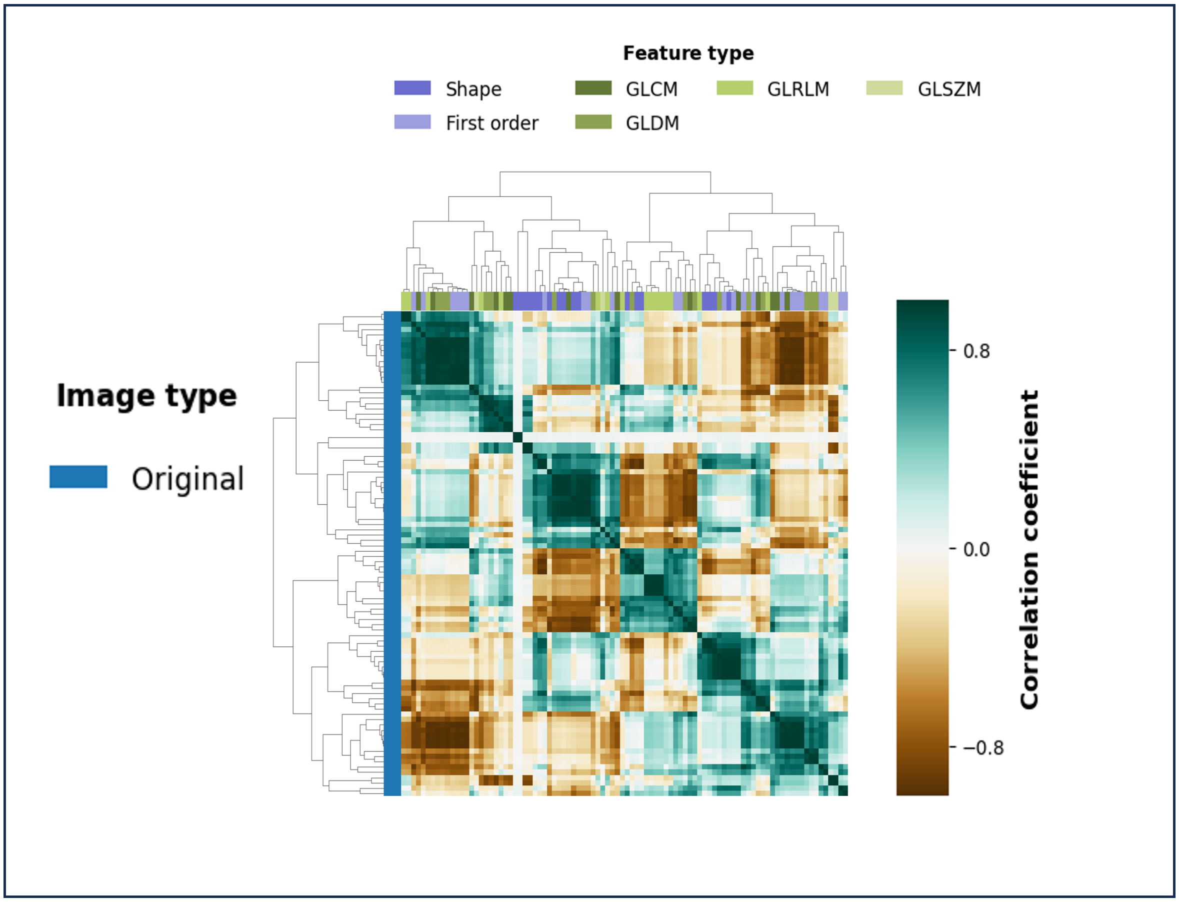



Figure 2.3: Hierarchically clustered heatmap of Pearson's correlation coefficients between the extracted 107 features (from original category only). Subsequent work will extract features from all image types to generate a larger feature set of 1373 texture features.

Forward SFS begins with an empty feature set and iteratively adds the feature that yields the highest performance gain at each step, continuing until no further improvement is observed [67, 68]. In contrast, backward SFS starts with the full feature set and progressively removes the feature whose exclusion improves model performance the most. Although forward selection is efficient, it evaluates each new feature only in the context of previously selected features, which can overlook feature interactions [69, 70]. Backward selection considers the contribution of each feature in the context of all other features and is often better at capturing such interactions. However, due to the extremely long computation time, only sequential feature selection in the forward direction was used to select 25 features.

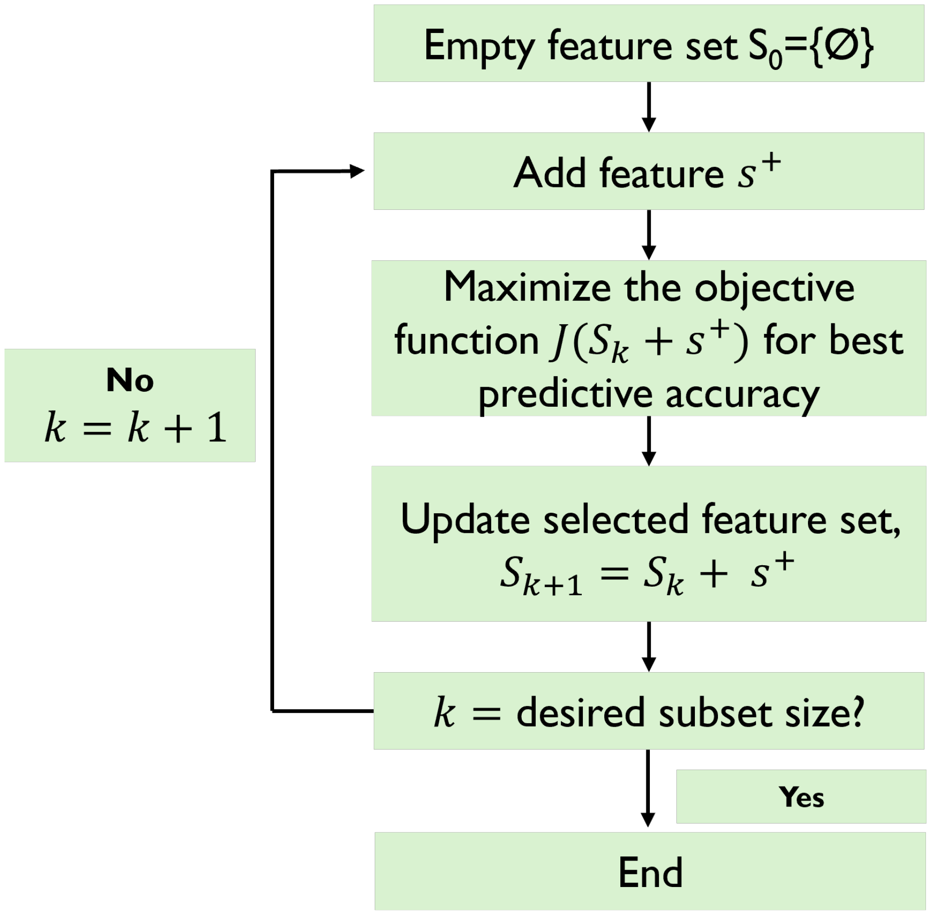
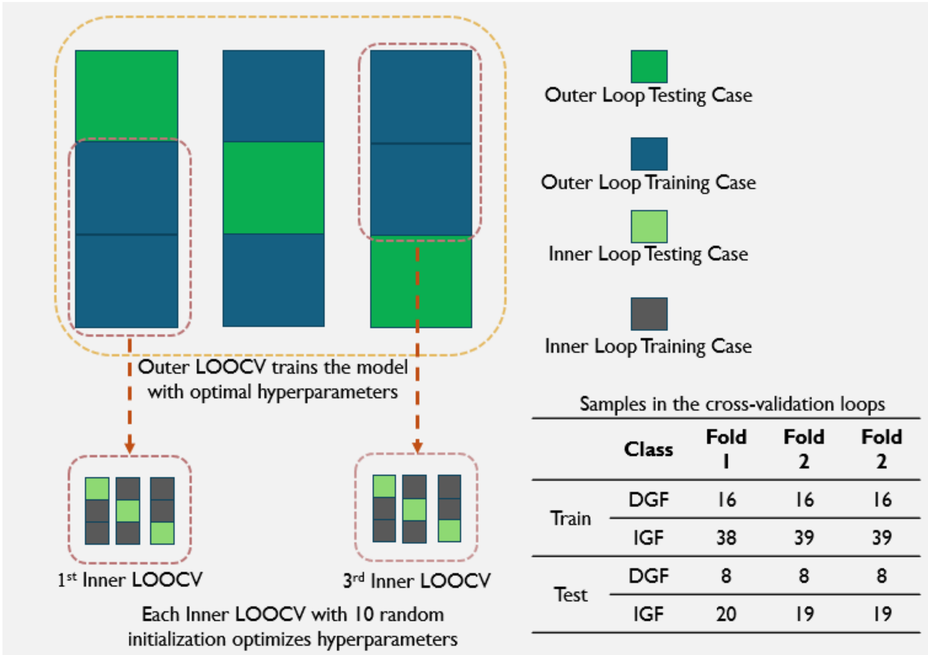



Figure 2.4: Sequential feature selection flowchart in the forward direction with five-fold cross-validation utilizing a Random Forest Regressor used to identify 25 (k = 25)optimal texture features.

The selected features were rescaled to 0-1 using minimum-maximum scaler prior to applying principal component analysis to reduce the data into lower dimensions. The variance threshold was set to 0.95 to identify features that are highly correlated and provide similar explanatory power, allowing retention of the most informative components while reducing redundancy [71].

## 2.3 Machine Learning Selection and Training

Five classifiers were selected for classifying delayed and immediate graft function (DGF and IGF respectively). These included decision tree (DT), support vector machines (SVM), extra trees (EXT) random forest (RF) and stochastic gradient descent (SGD). Each of these classifiers was has a class weight parameter that was used during cost-sensitive learning to give more weight to the minority class. A nested leave one-out cross validation (nested LOOCV, Figure 2.5) was used to train these classifiers. In the outer loop, the dataset was split into three folds with each fold being used only once during testing. In the inner loop, the other two folds are used to tune the model's hyperparameters. In the nested inner loop, the training data is also split into 3 folds (two for training and one for validation) in 10 different iterations providing 30 different configurations from which to select the optimal hyperparameters using GridSearch (Appendix A). The optimized model with the tuned hyperparameters was tested with the unseen testing data in the outer loop to generate the metrics used to compare various conditions (KDPI scores only, OCT texture data only and combination of both KDPI scores and OCT texture data). All analyses were performed on a Windows 11 operating system using Python version 3.12.




| | Class | Fold 1 | Fold 2 | Fold 2 |
|---|---|---|---|---|
| Train | DGF | 16 | 16 | 16 |
| | IGF | 38 | 39 | 39 |
| Test | DGF | 8 | 8 | 8 |
| | IGF | 20 | 19 | 19 |

Figure 2.5: Nested leave-one-out cross validation (LOOCV) with three folds in the inner and outer loops was used to during model training. The inner loop was used to tune the model's hyperparameters while the outer loop was used to for testing with unseen data.

## 2.4 Evaluation Metrics for Imbalanced Learning

Common evaluation metrics for predictive models such as accuracy and error rate are not designed for imbalanced data and emphasize the majority class, hence are misleading when used to evaluate classifiers [72, 73]. In binary classification, the evaluation metrics are derived from a confusion matrix of the actual versus predicted labels for the majority and minority class shown in Table 2.3. In this study, the receiver operating characteristic area (AUC ROC) curve, weighted precision, weighted recall, and geometric mean score were used to evaluate the performance of ML classifiers. Specificity (Equation 2.1, also known as precision or the true negative rate) evaluates the classifier's ability to predict the negative class (IGF). Similarly, sensitivity (Equation

2.2, also known as recall or true positive rate), evaluates the classifier's ability to predict the positive class (DGF). Both sensitivity and specificity only evaluate per-class accuracy of the model and do not evaluate whether the a classifier is biased towards one class or the other.

The geometric mean (Equation 2.3, g-mean) which takes the root of the product of sensitivity and recall was used to evaluate classifiers' bias towards either the negative and positive class. If the classifier is completely bias towards either class, the corresponding g-mean score would be zero. The higher the g-mean score, the lower the class bias. The AUC ROC is the area under the curve obtained by plotting sensitivity against 1-specificity. AUC ROC was used to compare the general ability of the classifiers to predict DGF and IGF.

Table 2.3: 2x2 Confusion Matrix for Binary Classification

| **Actual class** | **Predicted Class** | |
|---|---|---|
| | **DGF (-ve)** | **IGF (+ve)** |
| **DGF (-ve)** | True Positives (TP) | False Negative (FN) |
| **IGF (+ve)** | False Positive (FP) | True Negative (TN) |

$$\text{Specificity (precision)} = \frac{TN}{TP + FP} \tag{2.1}$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FP} \tag{2.2}$$

$$G\text{-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{2.3}$$

Cohen's Kappa (Equation2.4, k) was used to tune the classifiers' decision to account for data imbalance. This score evaluates the difference between a classifier's predicted labels and the actual labels. A Kappa score of 0 indicates that a classifier performed worse than a random untrained model, while a kappa score of 1 indicates that a classifier perfectly predicted the labels. The kappa score was used to identify the most optimal decision threshold that achieves a balance between specificity and accuracy, thereby, addressing the bias towards the majority class (IGF).

$$K = \frac{\text{accuracy} - \text{random accuracy}}{1 - \text{random accuracy}} \tag{2.4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

$$\text{Random accuracy} = \frac{(TP + FN)(TP + FP) + (TN + FP)(TN + FN)}{(TP + TN + FP + FN)^2} \tag{2.6}$$

# Chapter 3: Results

## 3.1 Optimizing the decision threshold

Imbalanced datasets are those with fewers in one category (minority class) compared to the other (majority class). Classifiers trained on imbalanced dataset tend to exhibit class bias towards the majority class. Common classifiers are designed with the assumption that the data sets are balanced and that misclassification costs are equally distributed among classes. However, in cases of class imbalance, classifiers tend to become biased towards the majority class. In this dataset, the majority class was immediate graft function (IGF, n = 65), significantly outnumbering the delayed graft function (DGF) minority class. To address this, I first implemented a method to adjust the classifier decision threshold from 0.5, followed by experiments using Synthetic Minority Oversampling with Tomek links (SMOTE-Tomek) to balance the data and implemented cost-sensitive learning.

The Generalized Heuristic for Optimizing Sensitivity Thresholds (GHOST) method was employed to automatically identify optimal decision thresholds for each classifier, addressing class imbalance by maximizing performance both sensitivity (recall) and specificity [45]. GHOST adjusts the classification decision threshold to achieve the best compromise between false positives and false negatives, thus improving the classification accuracy for minority classes. The GHOST algorithm was implemented in the outer-loop of the nested-cross validation after hyperparameter tuning using GridSearch.

It is worth noting that GHOST does not alter the classifier's accuracy because the area under the receiver operating curve remains the same. In binary classification, the geometric mean (g-mean) is the squared root of class-wise precision and recall. The geometric mean was used to evaluate the model's ability to classify both DGF and IGF effectively. Although optimizing the decision thresholds improved the geometric means of all the models trained on KDPI scores but SVM, the score plateau around 0.5 or below suggesting that the issue of class bias persisted (Table 3.1 and 3.2, Figure 3.1).

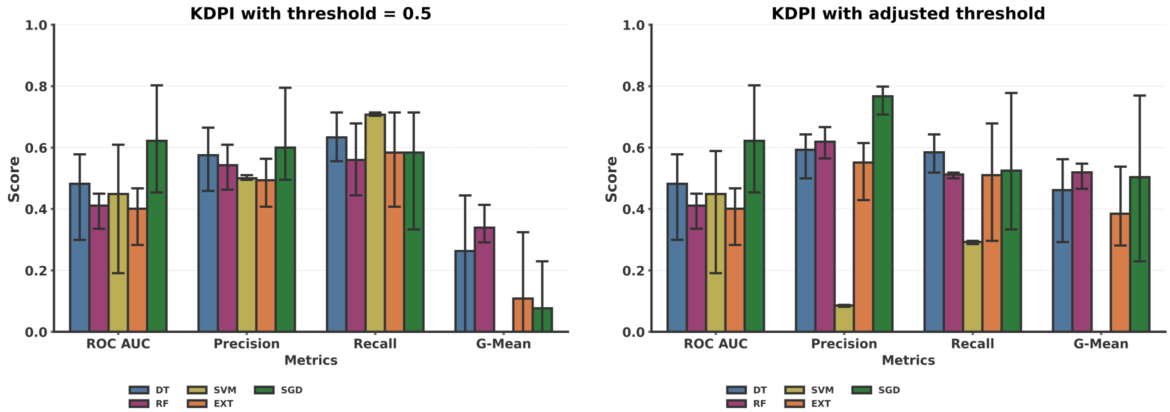
Table 3.1: Evaluation of classifiers trained on KDPI scores alone, t = 0.5

| **Models** | **ROC AUC** | **Precision** | **Recall** | **G-Mean** |
|---|---|---|---|---|
| DT | 0.4822 | 0.5751 | 0.6332 | 0.2630 |
| RF | 0.4110 | 0.5427 | 0.5595 | 0.3393 |
| SVM | 0.4487 | 0.5002 | 0.7072 | 0.0000 |
| EXT | 0.4010 | 0.4937 | 0.5838 | 0.1081 |
| SGD | 0.6220 | 0.6001 | 0.5838 | 0.0765 |

Table 3.2: Evaluation of classifiers trained on KDPI scores, $t = t_{optimal}$

| **Models** | **ROC AUC** | **Precision** | **Recall** | **G-Mean** |
|---|---|---|---|---|
| DT | 0.4822 | 0.5927 | 0.5847 | 0.4616 |
| RF | 0.4110 | 0.6193 | 0.5123 | 0.5193 |
| SVM | 0.4487 | 0.0857 | 0.2928 | 0.0000 |
| EXT | 0.4010 | 0.5512 | 0.5101 | 0.3848 |
| SGD | 0.6220 | 0.7670 | 0.5251 | 0.5037 |

A similar trend was observed for classifiers trained in OCT texture data only where all classifiers but SVM achieved a geometric means greater than 0.6 (Table 3.3 and 3.4, Figure 3.2). Therefore, adjusting the decision threshold improved the geometric means of the classifiers trained on both sets of data which implies that they had reduced class




(a) Default Threshold (t = 0.5) (b) Adjusted Threshold ($t = t_{optimal}$)

Figure 3.1: Comparison of classifiers trained on KDPI scores alone with the default and adjusted decision thresholds. DT: Decision Tree, RF: Random Forest,SVM: Support Vector Machine, EXT: Extra Trees, SGD: Stochastic Gradient Descent.

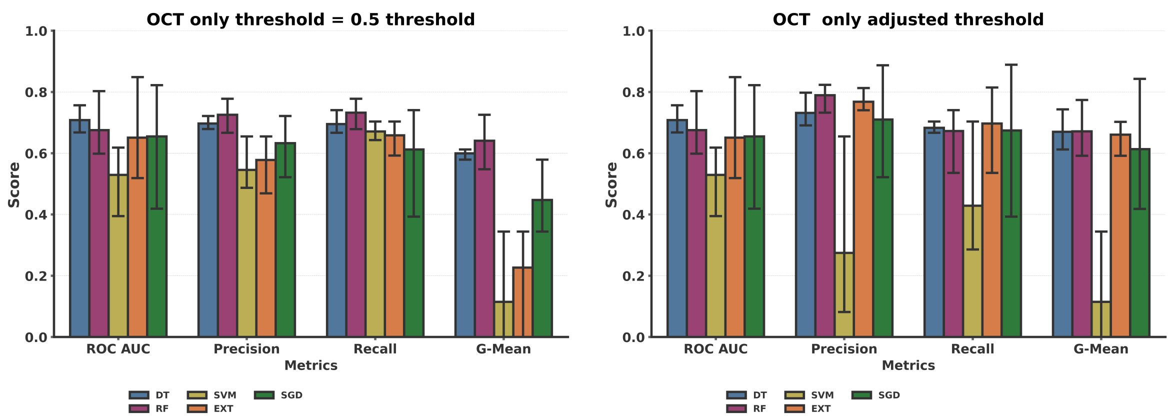
bias towards the majority category (IGF). The GHOST algorithm was used to adjust the decision thresholds in subsequent experiments.

Table 3.3: Evaluation of classifiers trained on OCT texture data only, $t = 0.5$

| **Models** | **ROC AUC** | **Precision** | **Recall** | **G-Mean** |
|---|---|---|---|---|
| DT | 0.7081 | 0.6972 | 0.6953 | 0.5995 |
| RF | 0.6754 | 0.7256 | 0.7324 | 0.6407 |
| SVM | 0.5294 | 0.5455 | 0.6711 | 0.1147 |
| EXT | 0.6510 | 0.5778 | 0.6583 | 0.2265 |
| SGD | 0.6549 | 0.6327 | 0.6124 | 0.4472 |

Table 3.4: Evaluation of classifiers trained on OCT texture data only, $t = t_{optimal}$

| Models | ROC AUC | Precision | Recall | G-Mean |
|---|---|---|---|---|
| DT | 0.7081 | 0.7317 | 0.6830 | 0.6699 |
| RF | 0.6754 | 0.7895 | 0.6724 | 0.6714 |
| SVM | 0.5294 | 0.2747 | 0.4286 | 0.1147 |
| EXT | 0.6510 | 0.7680 | 0.6971 | 0.6608 |
| SGD | 0.6549 | 0.7102 | 0.6742 | 0.6135 |



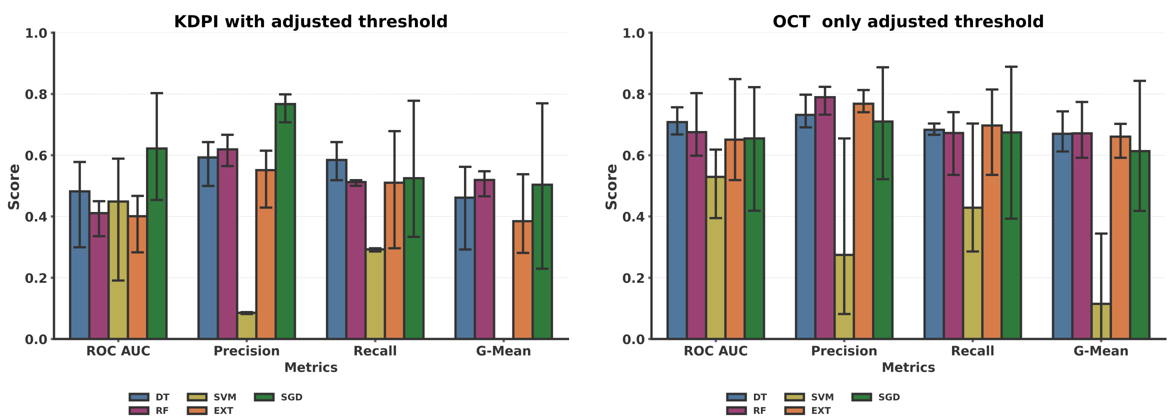

(a) Default Threshold (t = 0.5) (b) Adjusted Threshold ($t = t_{optimal}$)

Figure 3.2: Comparison of the performance of classifiers fitted with OCT texture data only with default and adjusted decision thresholds. DT: Decision Tree, RF: Random Forest,SVM: Support Vector Machine, EXT: Extra Trees, SGD: Stochastic Gradient Descent.

## 3.2 Baseline: OCT-texture versus KDPI scores

The Kidney Donor Profile Index (KDPI) is used to measure the suitability of deceased donor kidney for transplantation [74]. In this study, we first evaluated whether classifiers trained on KDPI scores alone can effectively discriminate between post-transplantation outcomes: immediate graft function (IGF, where the donor does not need dialysis one week after receiving renal transplantation) and delayed graft function (DGF, where the donor requires dialysis within one week after renal transplantation).

Five classifiers were trained using KDPI scores and post-transplantation outcomes of 87 diseased donor kidneys transplantations (DDKTs) of which 58 had IGF and 24 had DGF. Four of the five classifiers achieved a lower classification accuracy of below 50% (DT, RF, SVM, EXT) while SGD attained an accuracy of 62.20% (Table 3.2, Figure 3.3a). The accuracy metric evaluates the model's predictive ability to predict renal post-transplantation outcome compared to a random untrained model. The lower accuracy of DT, RF, SVM and EXT coupled with geometric mean scores below 0.5 for all models suggests that not only are classifiers trained on KDPI scores alone ineffective in predicting DGF/IGF status, but also exhibited a class bias towards the majority class (IGF), which has disproportionately more samples.




(a) KDPI scores only (t = $t_{optimal}$)

(b) OCT texture data only (t = $t_{optimal}$)

Figure 3.3: Comparison of classifiers fitted with KDPI scores alone (A) and OCT texture data alone (B). DT: Decision Tree, RF: Random Forest, SVM: Support Vector Machine, EXT: Extra Trees, SGD: Stochastic Gradient Descent.

Texture data was extracted using the PyRadiomics module from OCT images taken from kidney donors before transplantation and immediately after transplantation. Given that KDPI is only used to evaluate deceased donor kidney transplants, data from living donor kidney transplantations were excluded in this study. The original OCT images were used without any transformation in the texture feature extraction but subsequent work will include other image types (Table 2.2, Figure 2.3). Feature selection was done using sequential feature selection in the forward and backward direction to identify 25 most meaningful features for classifying DGF/IGF without KDPI scores. The selected features were used to train classifiers with or without KDPI scores. This data was further processed using kernel principal component analysis to identify optimal components (p = 5) that allow greater separability of the two classes and dimension reduction. The resulting data was used to train five classifiers to predict post-transplantation outcomes.

Compared to the same classifiers trained on KDPI scores alone, those trained using this OCT texture only achieved a higher prediction accuracy $\geq$60% and relatively higher geometric mean scores (Table 3.4, Figure 3.3b). The improvement in the performance of all the classifiers suggests that OCT derived texture data captures more information such as kidney microstructure that is not provided in the calculation of KDPI scores.

## 3.3 Combined OCT texture data and KDPI scores

When fitted with a combination of OCT texture data and KDPI scores, all classifiers but SVM underperformed in area under curve (ROC AUC) score compared to those trained on OCT texture data alone but still better than those trained on KDPI scores alone. Interestingly, classifiers trained with the combined dataset achieved comparable performance to those trained on OCT texture data alone but notably higher than those trained on KDPI scores only. The higher geometric means of the classifiers trained with OCT texture data only or the combined dataset suggests that they have lower class bias

Table 3.5: Evaluation of classifiers combined OCT texture and KDPI data

| Models | ROC AUC | Precision | Recall | G-Mean |
|---|---|---|---|---|
| DT | 0.5845 | 0.6776 | 0.6706 | 0.5310 |
| RF | 0.6417 | 0.7202 | 0.7196 | 0.6361 |
| SVM | 0.5607 | 0.6478 | 0.6830 | 0.4288 |
| EXT | 0.5986 | 0.7127 | 0.6834 | 0.6467 |
| SGD | 0.5679 | 0.6422 | 0.6583 | 0.5107 |

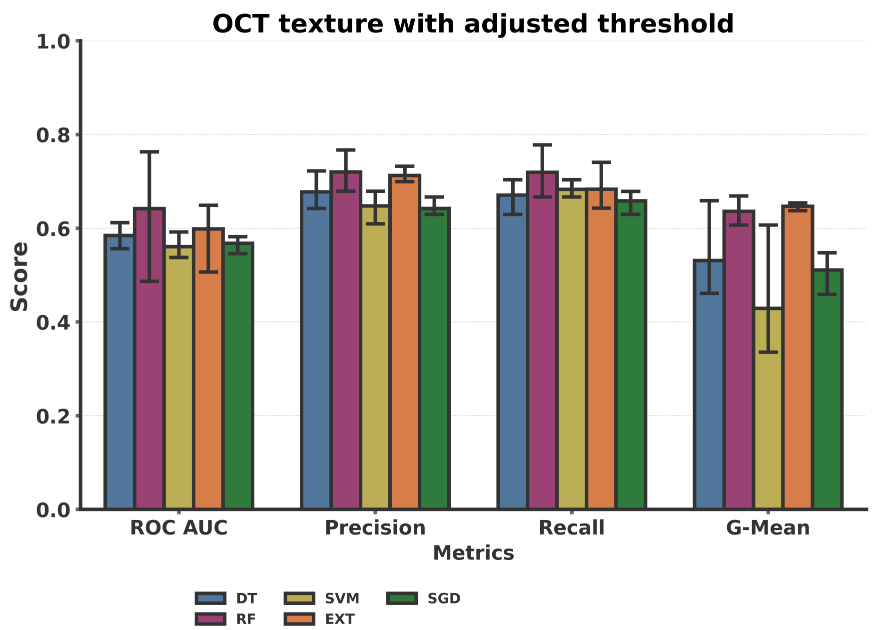



Figure 3.4: Comparison of the performance of classifiers fitted with a combination of the OCT-derived texture data and KDPI scores. DT: Decision Tree, RF: Random Forest,SVM: Support Vector Machine, EXT: Extra Trees, SGD: Stochastic Gradient Descent.

towards the majority class (IGF). With OCT data, the classifiers are able to effectively classify both delayed (DGF) and immediate graft function (IGF).

## 3.4 Balancing the data with SMOTE-Tomek

Following threshold optimization, I tested the effect of balancing the combined dataset with a variant of Synthetic Minority Oversampling which includes Tomek-links. SMOTE is a method that combines both oversampling of the minority class using synthetic data and under sampling the minority class [48]. In the case of SMOTE-Tomek, SMOTE generated synthetic minority (DGF) samples while Tomek links are used to clear majority class samples from class boundaries to improve classifier's performance [48, 75].

While SMOTE-Tomek reduced the area under curve (ROC AUC) across all models, SGD exhibited mixed effects, showing variable performance in their geometric mean scores (Table 3.6, Figure 3.5). Stochastic Gradient Descent (SGD) outperformed the other classifiers and had an improvement in both the accuracy and g-mean scores suggesting that balancing the imbalanced dataset with SMOTE-Tomek can be an alternative solution. Interestingly, [76] found that SGD consistently outperformed 18 other classifiers in tumor type prediction across multiple experiments involving imbalanced data resampling techniques, including SMOTE. Although SGD is sensitive to feature scaling and requires extensive hyperparameter tuning of the regularization parameter and the number of iterations, it is simple yet effective at large scale learning problems [77, 78].

Table 3.6: Evaluation of classifiers after balancing data using SMOTE-Tomek

| **Models** | **ROC AUC** | **Precision** | **Recall** | **G-Mean** |
|---|---|---|---|---|
| 0.5497 | 0.6446 | 0.5608 | 0.5528 | 0.6272 |
| 0.5717 | 0.6827 | 0.6455 | 0.5981 | 0.6722 |
| 0.5559 | 0.6746 | 0.5631 | 0.5512 | 0.3590 |
| 0.4608 | 0.6594 | 0.5503 | 0.5382 | 0.6215 |
| 0.6271 | 0.7734 | 0.6226 | 0.6473 | 0.5832 |

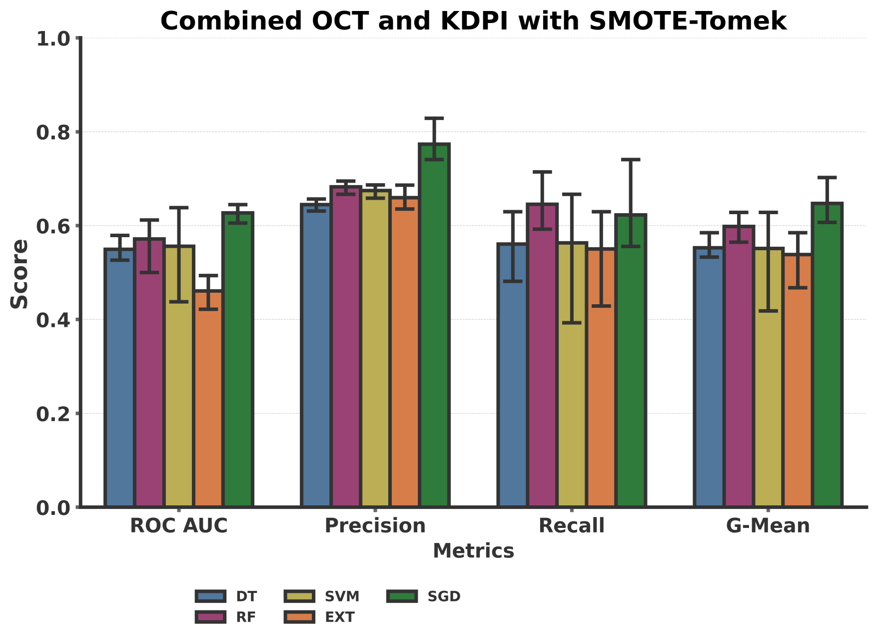



Figure 3.5: Comparison of the performance of classifiers fitted with a combination of the OCT-derived texture data and KDPI scores after balancing it using SMOTE-Tomek. SMOTE-Tomek: Synthetic Minority Oversampling Technique with Tomek links, DT: Decision Tree, RF: Random Forest,SVM: Support Vector Machine, EXT: Extra Trees, SGD: Stochastic Gradient Descent.

## 3.5 Cost-sensitive learning

Lastly, I implemented cost-sensitive learning (CSL) to give higher weight to the minority class (DGF). The imbalanced ratio (IR), defined as the ratio between the samples in the minority class and the majority class was used to adjust the class-weight parameter to give more weight to the DGF category. The class weight for the IGF category was set to the IR ratio, while that of the DGF category was set to 1.

Compared to classifiers fitted with the combined OCT texture and KDPI dataset, implementing cost-sensitive learning (CSL) improved both area under curve (AUC) and geometric means for the DT, RF, and SGD models, but led to reduced performance for the EXT and SVM models compared to same classifiers trained without CSL (Table 3.7,

Figure 3.6). As anticipated, implementing CSL resulted in relatively higher geometric means but only for DT, RF, and SVM while EXT and SGD had comparable scores to those trained with the combined OCT texture and KDPI scores only.

Table 3.7: Evaluation of classifiers after implementing Cost-Sensitive Learning

| Models | ROC AUC | Precision | Recall | G-Mean |
|---|---|---|---|---|
| DT | 0.6114 | 0.6918 | 0.6332 | 0.6272 |
| RF | 0.6526 | 0.7301 | 0.7191 | 0.6722 |
| SVM | 0.5281 | 0.6889 | 0.5966 | 0.3590 |
| EXT | 0.5279 | 0.6924 | 0.6102 | 0.6215 |
| SGD | 0.6003 | 0.7121 | 0.5494 | 0.5832 |

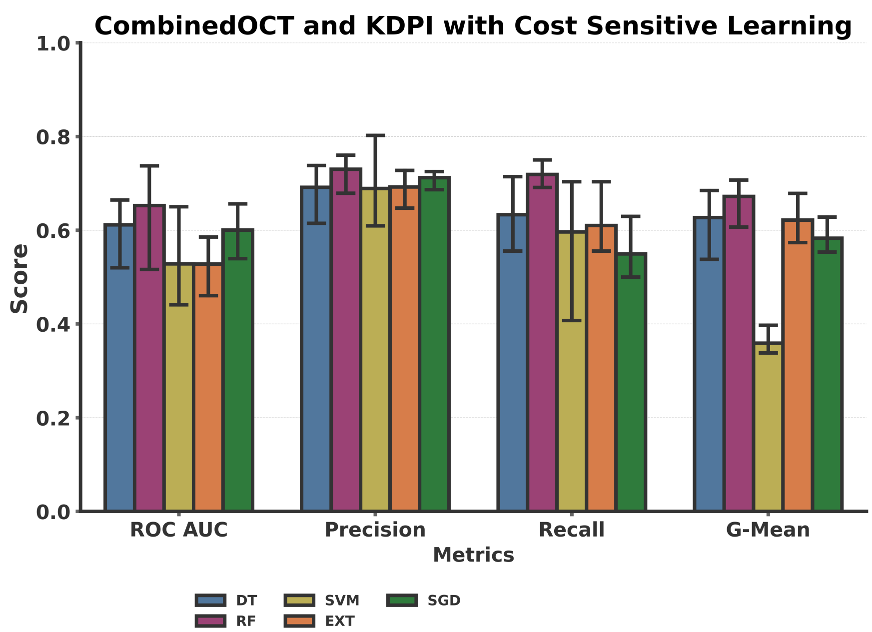



Figure 3.6: Comparison of the performance of classifiers fitted with a combination of the OCT-derived texture data and KDPI scores after implementing cost-sensitive learning to give higher weight to the minority (DGF) category. SMOTE-Tomek: Synthetic Minority Oversampling Technique with Tomek links, DT: Decision Tree, RF: Random Forest,SVM: Support Vector Machine, EXT: Extra Trees, SGD: Stochastic Gradient Descent.

# Chapter 4: Conclusion and Future Work

The primary goal of this study was to evaluate whether classifiers trained on kidney donor profile index (KDPI) scores alone were effective at classifying delayed graft function (DGF) from immediate graft function (IGF). To this end, I compared the performance of five classifiers (decision tree, DT; support vector machines, SVM; extra trees, EXT; random forest, RF; and stochastic gradient descent, SGD) fitted with either KDPI scores alone, OCT texture data only or a combined dataset comprising of both OCT texture data and KDPI scores.

Classifiers fitted with OCT texture data only performed higher than both those fitted with KDPI scores alone and combined dataset. Similarly, those trained with the combined dataset outperformed the KDPI only condition which implies that KDPI scores are ineffective at predicting the post-transplantation outcomes (DGF and IGF). A potential confounding aspect of this study involves the overreliance on KDPI scores which are not meant predict post-transplantation outcomes but rather, to assess the longevity of a deceased donor kidney relative to all the kidneys from the previous years. However, these findings suggest that the texture data extracted from the optical coherence tomography (OCT) images captured meaningful features that are useful for predicting the post-transplantation outcomes.

Our previous work using OCT images focused on characterizing the tubules (proximal convoluted tubules and distal convoluted tubules) that are visible in kidney OCT images [24]. The limitation of that study was focusing on a single parameter (tubular information) as opposed to the current study's use of a variety of texture data. However, future work will evaluate other parameters from the donor or recipient history that can be useful for predicting the post-transplantation outcomes.

The dataset used in this study only comprised of deceased donors with disproportionately more patients with immediate graft function (n = 58) than delayed graft (n = 24) function. The ratio between the delayed and immediate graft function patients represents the imbalanced ratio (IR). Since common classifiers are designed under the assumption of even samples between the categories, it was necessary to implement solutions to address their bias towards the majority samples. In addition to precision and recall which are class-wise accuracy metrics, I used the geometric mean which is the squared root of the product of precision and recall to evaluate whether the classifiers were biased towards IGF or they were effectively classifying the minority class (DGF) as well.

During model training, I first implemented an iterative algorithm that balances the ratio of false positives and false negatives to tune the decision threshold instead of using the default threshold of 0.5. While this did not lower the area under the receiver operating characteristic curve (ROC AUC), it resulted in balanced scores for precision and recall and higher geometric means. This configuration, which I termed adjusted threshold, was used for the subsequent comparisons. I also implemented synthetic minority oversampling technique to upsample the minority class by creating synthetic data and used Tomek-links to undersample the majority class by removing majority samples at the decision margin.

Another solution I tested was cost-sensitive learning (CSL) where I gave higher weight to the minority class to address class bias. Both SMOTE-Tomek and CSL yielded mixed results in accuracy but with nearly all models having higher geometric means. While these solutions did not increase the overall predictive performance of the classifiers, they resulted in stable performance across models and lower bias towards the majority class (IGF). These mixed results also suggest that a each classifiers would benefit from a tailored solution as opposed to a one-size-fits-all solution during imbalanced data learning.

Future work on this project will address various limitations of this study. First, I will evaluate whether the capsule, the outer layer of the kidney is essential for predicting post-transplantation function. This layer is the brightest on the OCT images and might skew the texture data from the inner tissues. I will also investigate whether other factors from the donor and recipient history can outperform KDPI in the combined dataset. The nested cross-validation used in this benchmark study comprised of three folds but future work will utilize four folds to ensure sufficient training data given the few patients in the minority class. The OCT images used in this study were scanned from various locations in the kidney but those locations were not tracked. Our collaborators are actively collecting new data but tracking the scanning location with a standardized protocol which will allow us to predict which parts of the kidney are predisposed to delayed versus immediate graft function outcomes.

# References


[1] Krista L. Lentine, Jodi M. Smith, Grace R. Lyden, Jonathan M. Miller, Sarah E. Booker, Thomas G. Dolan, Kayla R. Temple, Samantha Weiss, Dzhuliyana Handarova, Ajay K. Israni, and Jon J. Snyder. OPTN/SRTR 2023 Annual Data Report: Kidney. *American Journal of Transplantation*, 25(2, Supplement 1):S22–S137, 2025.

[2] CDC. Chronic Kidney Disease in the United States, 2023, May 2024.

[3] K. J. Ko, Y. H. Kim, K. H. Kwon, M. H. Kim, K. W. Jun, J. K. Hwang, S. D. Kim, S. C. Park, J. I. Kim, S. S. Yun, and I. S. Moon. Kidney Transplantation Using Expanded-Criteria Deceased Donors: A Comparison With Ideal Deceased Donors and Non-Expanded-Criteria Deceased Donors. *Transplantation Proceedings*, 50(10):3222–3227, December 2018.

[4] Johan Noble, Thomas Jouve, Paolo Malvezzi, Caner Süsal, and Lionel Rostaing. Transplantation of Marginal Organs: Immunological Aspects and Therapeutic Perspectives in Kidney Transplantation. *Frontiers in Immunology*, 10, January 2020. Publisher: Frontiers.

[5] Qasem Al-Tashi, Maliazurina B. Saad, Amgad Muneer, Rizwan Qureshi, Seyedali Mirjalili, Ajay Sheshadri, Xiuning Le, Natalie I. Vokes, Jianjun Zhang, and Jia Wu. Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review. *International Journal of Molecular Sciences*, 24(9):7781, April 2023.

[6] Julio Pascual, Javier Zamora, and John D. Pirsch. A Systematic Review of Kidney Transplantation From Expanded Criteria Donors. *American Journal of Kidney Diseases*, 52(3):553–586, September 2008.

[7] Sri G. Yarlagadda, Steven G. Coca, Amit X. Garg, Mona Doshi, Emilio Poggio, Richard J. Marcus, and Chirag R. Parikh. Marked variation in the definition and diagnosis of delayed graft function: a systematic review. *Nephrology Dialysis Transplantation*, 23(9):2995–3003, September 2008.

[8] Dermot H. Mallon, Dominic M. Summers, J. Andrew Bradley, and Gavin J. Pettigrew. Defining Delayed Graft Function after Renal Transplantation: Simplest Is Best. *Transplantation*, 96(10):885, November 2013.

[9] A. Siedlecki, W. Irish, and D. C. Brennan. Delayed Graft Function in the Kidney Transplant. *American Journal of Transplantation*, 11(11):2279–2296, November 2011.

[10] F. Catena, F. Coccolini, G. Montori, C. Vallicelli, A. Amaduzzi, G. Ercolani, M. Ravaioli, M. Del Gaudio, R. Schiavina, E. Brunocilla, G. Liviano, G. Feliciangeli, and A. D. Pinna. Kidney Preservation: Review of Present and Future Perspective. *Transplantation Proceedings*, 45(9):3170–3177, November 2013.

[11] Charles Y Lee and Martin J Mangino. Preservation methods for kidney and liver. *Organogenesis*, 5(3):105–112, 2009.

[12] Sarah A. Hosgood, Rachel J. Brown, and Michael L. Nicholson. Advances in Kidney Preservation Techniques and Their Application in Clinical Practice. *Transplantation*, 105(11):e202, November 2021.

[13] Ina Jochmans, John M. O'Callaghan, Jacques Pirenne, and Rutger J. Ploeg. Hypothermic machine perfusion of kidneys retrieved from standard and high-risk donors. *Transplant International*, 28(6):665–676, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tri.12530.

[14] Peri Husen, Catherine Boffa, Ina Jochmans, Christina Krikke, Lucy Davies, Laura Mazilescu, Aukje Brat, Simon Knight, Daniel Wettstein, Orsolya Cseprekal, Neal Banga, Maria Irene Bellini, Laszlo Szabo, Elijah Ablorsu, Tom Darius, Isabel Quiroga, Michel Mourad, Johann Pratschke, Vassilios Papalois, Zoltan Mathe, Henri G. D. Leuvenink, Thomas Minor, Jacques Pirenne, Rutger J. Ploeg, and Andreas Paul. Oxygenated End-Hypothermic Machine Perfusion in Expanded Criteria Donor Kidney Transplant. *JAMA Surgery*, 156(6):1–9, June 2021.

[15] Aukje Brat, Kirsten M. de Vries, Ernst W. E. van Heurn, Volkert A. L. Huurman, Wim de Jongh, Henri G. D. Leuvenink, Arjan D. van Zuilen, Bernadette J. J. M. Haase-Kromwijk, Jeroen de Jonge, Stefan P. Berger, and Sijbrand H. Hofker. Hypothermic Machine Perfusion as a National Standard Preservation Method for Deceased Donor Kidneys. *Transplantation*, 106(5):1043, May 2022.

[16] Feng Yan, Bornface Mutembei, Trisha Valerio, Gokhan Gunay, Ji-Hee Ha, Qinghao Zhang, Chen Wang, Ebenezer Raj Selvaraj Mercyshalinie, Zaid A. Alhajeri, Fan Zhang, Lauren E. Dockery, Xinwei Li, Ronghao Liu, Danny N. Dhanasekaran, Handan Acar, Wei R. Chen, and Qinggong Tang. Optical coherence tomography for multicellular tumor spheroid category recognition and drug screening classification via multi-spatial-superficial-parameter and machine learning. *Biomedical Optics Express*, 15(4):2014–2047, April 2024. Publisher: Optica Publishing Group.

[17] Yongyang Huang, Shunqiang Wang, Qiongyu Guo, Sarah Kessel, Ian Rubinoff, Leo Li-Ying Chan, Peter Li, Yaling Liu, Jean Qiu, and Chao Zhou. Optical Coherence Tomography Detects Necrotic Regions and Volumetrically Quantifies Multicellular Tumor Spheroids. *Cancer Research*, 77(21):6011–6020, October 2017.

[18] James G. Fujimoto. Optical coherence tomography for ultrahigh resolution in vivo imaging. *Nature Biotechnology*, 21(11):1361–1367, November 2003. Publisher: Nature Publishing Group.

[19] Qian Li, Maristela L. Onozato, Peter M. Andrews, Chao-Wei Chen, Andrew Paek, Renee Naphas, Shuai Yuan, James Jiang, Alex Cable, and Yu Chen. Automated quantification of microstructural dimensions of the human kidney using optical coherence tomography (OCT). *Optics Express*, 17(18):16000–16016, August 2009. Publisher: Optica Publishing Group.

[20] Yu Chen, Peter M. Andrews, Aaron D. Aguirre, Joseph M. Schmitt, and James G. Fujimoto. High-resolution three-dimensional optical coherence tomography imaging of kidney microanatomy ex vivo. *Journal of Biomedical Optics*, 12(3):034008, May 2007. Publisher: SPIE.

[21] Xihan Ma, Mousa Moradi, Xiaoyu Ma, Qinggong Tang, Moshe Levi, Yu Chen, and Haichong K. Zhang. Large Area Kidney Imaging for Pre-transplant Evaluation using Real-Time Robotic Optical Coherence Tomography. *Research Square*, pages rs.3.rs–3385622, October 2023.

[22] Peter M Andrews, Yu Chen, Maristela L Onozato, Shu-Wei Huang, Desmond C Adler, Robert A Huber, James Jiang, Scott E Barry, Alex E Cable, and James G Fujimoto. High-resolution optical coherence tomography imaging of the living kidney. *Laboratory Investigation*, 88(4):441–449, April 2008.

[23] Peter M. Andrews, Yu Chen, Jeremiah Wierwille, Daniel Joh, Peter Alexandrov, Derek Rogalsky, Patrick Moody, Allen Chen, Matthew Cooper, Jennifer E. Verbesey, Wei Gong, and Hsing-Wen Wang. Using OCT to predict post-transplant renal function. In *Photonic Therapeutics and Diagnostics IX*, volume 8565, pages 156–163. SPIE, March 2013.

[24] Brandon Konkel, Christopher Lavin, Tong Tong Wu, Erik Anderson, Aya Iwamoto, Hadi Rashid, Brandon Gaitian, Joseph Boone, Matthew Cooper, Peter Abrams, Alexander Gilbert, Qinggong Tang, Moshe Levi, James G. Fujimoto, Peter Andrews, and Yu Chen. Fully automated analysis of OCT imaging of human kidneys for prediction of post-transplant function. *Biomedical Optics Express*, 10(4):1794–1821, March 2019.

[25] Arvid B. Maunsbach. The influence of different fixatives and fixation methods on the ultrastructure of rat kidney proximal tubule cells: I. Comparison of different perfusion fixation methods and of glutaraldehyde, formaldehyde and osmium tetroxide fixatives. *Journal of Ultrastructure Research*, 15(3):242–282, June 1966.

[26] C. J. Wang, J. B. Wetmore, G. S. Crary, and B. L. Kasiske. The Donor Kidney Biopsy and Its Implications in Predicting Graft Outcomes: A Systematic Review. *American Journal of Transplantation*, 15(7):1903–1914, July 2015.

[27] Mark Haidekker. Texture Analysis. In *Advanced Biomedical Image Analysis*, pages 236–275. IEEE, 2011.

[28] Maria Petrou. Texture in Biomedical Images. In Thomas Martin Deserno, editor, *Biomedical Image Processing*, pages 157–176. Springer, Berlin, Heidelberg, 2011.

[29] Adrien Depeursinge, Omar S Al-Kadi, and J.Ross Mitchell. *Biomedical Texture Analysis*. Elsevier/Academic Press, 1 edition, August 2017.

[30] Zhenyu Liu, Shuo Wang, Di Dong, Jingwei Wei, Cheng Fang, Xuezhi Zhou, Kai Sun, Longfei Li, Bo Li, Meiyun Wang, and Jie Tian. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. *Theranostics*, 9(5):1303–1322, February 2019.

[31] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G. P. M. van Stiphout, Patrick Granton, Catharina M. L. Zegers, Robert Gillies, Ronald Boellard, André Dekker, and Hugo J. W. L. Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, March 2012.

[32] A. Bizzego, N. Bussola, D. Salvalai, M. Chierici, V. Maggio, G. Jurman, and C. Furlanello. Integrating deep and radiomics features in cancer bioimaging. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8, July 2019.

[33] Fanny Orlhac, Christophe Nioche, Michaël Soussan, and Irène Buvat. Understanding Changes in Tumor Texture Indices in PET: A Comparison Between Visual Assessment and Index Values in Simulated and Patient Data. *Journal of Nuclear Medicine*, 58(3):387–392, March 2017. Publisher: Society of Nuclear Medicine Section: Oncology.

[34] Anne Humeau-Heurtier. Texture Feature Extraction Methods: A Survey. *IEEE Access*, 7:8975–9000, 2019.

[35] Ming Zheng, Tong Li, Xiaoyao Zheng, Qingying Yu, Chuanming Chen, Ding Zhou, Changlong Lv, and Weiyi Yang. UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification. *Information Sciences*, 576:658–680, October 2021.

[36] Mabrouka Salmi, Dalia Atif, Diego Oliva, Ajith Abraham, and Sebastian Ventura. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57(10):273, September 2024.

[37] Karoline Kernig, Veronica Albrecht, Desirée-Louise Dräger, Andreas Führer, Steffen Mitzner, Günther Kundt, and Oliver W. Hakenberg. Predictors of Delayed Graft Function in Renal Transplantation. *Urologia Internationalis*, 106(5):512–517, December 2021.

[38] Wanwan Zheng and Mingzhe Jin. The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study. *SN Computer Science*, 1(2):71, February 2020.

[39] YANMIN SUN, ANDREW K. C. WONG, and MOHAMED S. KAMEL. CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, November 2011. Publisher: World Scientific Publishing Company.

[40] Shaza M. Abd Elrahman and Ajith Abraham. A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing*, 1:9–9, October 2013.

[41] Vitor Werner de Vargas, Jorge Arthur Schneider Aranda, Ricardo dos Santos Costa, Paulo Ricardo da Silva Pereira, and Jorge Luis Victória Barbosa. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, 65(1):31–57, 2023.

[42] Ginny Y. Wong, Frank H.F. Leung, and Sai-Ho Ling. A novel evolutionary preprocessing method based on over-sampling and under-sampling for imbalanced datasets. In *IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society*, pages 2354–2359, November 2013. ISSN: 1553-572X.

[43] Elaheh Jafarigol and Theodore B. Trafalis. A distributed approach to meteorological predictions: addressing data imbalance in precipitation prediction models through federated learning and GANs. *Computational Management Science*, 21(1):22, February 2024.

[44] Victor Henrique Alves Ribeiro and Gilberto Reynoso-Meza. Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. *Expert Systems with Applications*, 147:113232, June 2020.

[45] Carmen Esposito, Gregory A. Landrum, Nadine Schneider, Nikolaus Stiefl, and Sereina Riniker. GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling*, 61(6):2623–2640, June 2021.

[46] Debashree Devi, Saroj K. Biswas, and Biswajit Purkayastha. A review on solution to class imbalance problem: Undersampling approaches. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 626–631, 2020.

[47] Vinod Kumar, Gotam Singh Lalotra, Ponnusamy Sasikala, Dharmendra Singh Rajput, Rajesh Kaluri, Kuruva Lakshmanna, Mohammad Shorfuzzaman, Abdulmajeed Alsufyani, and Mueen Uddin. Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare*, 10(7):1293, July 2022.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, June 2011. arXiv:1106.1813.

[49] Alberto Fernandez, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, April 2018.

[50] Gustavo E. A. P. A. Batista, Ana L. C. Bazzan, and Maria Carolina Monard. Balancing Training Data for Automated Annotation of Keywords: a Case Study.

[51] Anuraganand Sharma, Prabhat Kumar Singh, and Rohitash Chandra. SMOTified-GAN for Class Imbalanced Pattern Classification Problems. *IEEE Access*, 10:30655–30665, 2022. Conference Name: IEEE Access.

[52] Ivan Tomek. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, November 1976. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

[53] Claudio Jeldres, Héloïse Cardinal, Alain Duclos, Shahrokh F. Shariat, Nazareno Suardi, Umberto Capitanio, Marie-Josèe Hébert, and Pierre I. Karakiewicz. Prediction of delayed graft function after renal transplantation. *Canadian Urological Association Journal*, 3(5):377–382, October 2009.

[54] Alexander Decruyenaere, Philippe Decruyenaere, Patrick Peeters, Frank Vermassen, Tom Dhaene, and Ivo Couckuyt. Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods. *BMC Medical Informatics and Decision Making*, 15(1):83, October 2015.

[55] Gülşah Kaya Aksoy, Hüseyin Gökhan Akçay, Çağlar Arı, Mehtap Adar, Mustafa Koyun, Elif Çomak, and Sema Akman. Predicting graft survival in paediatric kidney transplant recipients using machine learning. *Pediatric Nephrology*, 40(1):203–211, January 2025.

[56] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks, November 2018. arXiv:1611.07004 [cs].

[57] Joyce Henry, Terry Natalie, and Den Madsen. Pix2pix gan for image-to-image translation. *Research Gate Publication*, pages 1–5, 2021.

[58] Hong-an Li, Min Zhang, Zhenhua Yu, Zhanli Li, and Na Li. An improved pix2pix model based on gabor filter for robust color image rendering. *Math. Biosci. Eng*, 19(1):86–101, 2022.

[59] Luca Tirel, Ali Mohamed Ali, and Hashim A Hashim. Novel hybrid integrated pix2pix and wgan model with gradient penalty for binary images denoising. *Systems and Soft Computing*, 6:200122, 2024.

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].

[61] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024. Publisher: Nature Publishing Group.

[62] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[63] Leland Wilkinson, , and Michael Friendly. The History of the Cluster Heat Map. *The American Statistician*, 63(2):179–184, May 2009. Publisher: ASA Website _eprint: https://doi.org/10.1198/tas.2009.0033.

[64] Ke Zhang, Neman Abdoli, Patrik Gilley, Youkabed Sadri, Xuxin Chen, Theresa C. Thai, Lauren Dockery, Kathleen Moore, Robert S. Mannel, and Yuchen Qiu. Developing a Novel Image Marker to Predict the Clinical Outcome of Neoadjuvant Chemotherapy (NACT) for Ovarian Cancer Patients. *ArXiv*, page arXiv:2309.07087v2, July 2024.

[65] Xue-wen Chen and Jong Cheol Jeong. Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 429–435, December 2007.

[66] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422, January 2002.

[67] David W. Aha and Richard L. Bankert. A Comparative Evaluation of Sequential Feature Selection Algorithms. In Doug Fisher and Hans-J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 199–206. Springer, New York, NY, 1996.

[68] Joseph Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn. Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition. *Biomolecules*, 10(3):454, March 2020. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[69] Huan Liu and Hiroshi Motoda. Feature Selection Methods. In Huan Liu and Hiroshi Motoda, editors, *Feature Selection for Knowledge Discovery and Data Mining*, pages 73–95. Springer US, Boston, MA, 1998.

[70] Justin Doak. An Evaluation of Feature Selection Methods.

[71] Harun Uğuz. A Biomedical System Based on Artificial Neural Network and Principal Component Analysis for Diagnosis of the Heart Valve Diseases. *Journal of Medical Systems*, 36(1):61–72, February 2012.

[72] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3(10):27, 2013.

[73] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251, September 2013. ISSN: 2156-8111.

[74] OPTN. A Guide to Calculating and Interpreting the Kidney Donor Profle Index (KDPI). 2020.

[75] Gelin Wang, Weiduo Zhao, Jiqiang Wang, Xinmin Chen, and Jing Li. A SMOTE-Tomek-Based Parameter Identification and Behavior Estimation Method for IPMSM in Aerial Applications. In Yingmin Jia, Weicun Zhang, Yongling Fu, and Huihua Yang, editors, *Proceedings of 2024 Chinese Intelligent Systems Conference*, pages 27–35, Singapore, 2024. Springer Nature.

[76] Yuting Yang and Golrokh Mirzaei. Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification. *PLOS ONE*, 19(2):e0293607, February 2024.

[77] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, page 116, New York, NY, USA, July 2004. Association for Computing Machinery.

[78] scikit-learn. 1.5. Stochastic Gradient Descent.

# 1 Appendix A

Table A.1: Hyperparameter tuning using GridSearch

| **Model** | **Hyperparameter** | **Values Tested** |
|---|---|---|
| Decision Tree | max_depth | [3, 5, 10, None] |
| | min_samples_split | [2, 5, 10] |
| | min_samples_leaf | [1, 2, 4] |
| Random Forest | n_estimators | [50, 100, 200] |
| | max_depth | [3, 5, 10, None] |
| | min_samples_split | [2, 5, 10] |
| SVM | C | [0.00001, 0.001, 0.01, 0.1, 1, 10, 100, 1000] |
| | kernel | ['linear', 'rbf', 'poly'] |
| | gamma | ['scale', 'auto'] |
| Extra Trees | n_estimators | [50, 100, 200] |
| | max_depth | [3, 5, 10, None] |
| | min_samples_split | [2, 5, 10] |
| SGD | alpha | [0.0001, 0.001, 0.01] |
| | max_iter | [1000, 2000] |