Machine Learning Algorithm Improves the Prediction of Transplant Hepatic Artery Stenosis or Occlusion

A Single-Center Study

Keith Feldman, PhD,† Justin Baraboo, MSc,‡ Deeyendal Dinakarpandian, PhD,§ and Sherwin S. Chan, MD, PhD‖¶*

Abstract: The aim of this study was to determine if machine learning can improve the specificity of detecting transplant hepatic artery pathology over conventional quantitative measures while maintaining a high sensitivity.

This study presents a retrospective review of 129 patients with transplanted hepatic arteries. We illustrate how beyond common clinical metrics such as stenosis and resistive index, a more comprehensive set of waveform data (including flow half-lives and Fourier transformed waveforms) can be integrated into machine learning models to obtain more accurate screening of stenosis and occlusion. We present a novel framework of Extremely Randomized Trees and Shapley values, we allow for explainability at the individual level.

The proposed framework identified cases of clinically significant stenosis and occlusion in hepatic arteries with a state-of-the-art specificity of 65%, while maintaining sensitivity at the current standard of 94%. Moreover, through 3 case studies of correct and mispredictions, we demonstrate examples of how specific features can be elucidated to aid in interpreting driving factors in a prediction.

This work demonstrated that by utilizing a more complete set of waveform data and machine learning methodologies, it is possible to reduce the rate of false-positive results in using ultrasounds to screen for transplant hepatic artery pathology compared with conventional quantitative measures. An advantage of such techniques is explainability measures at the patient level, which allow for increased radiologists' confidence in the predictions.

Key Words: machine learning, spectral Doppler ultrasound, liver transplant, transplant hepatic arterial stenosis.

(*Ultrasound Quarterly* 2023;39: 86–94)

Each year, more than 9000 Americans will undergo liver transplant because of conditions such as acute liver failure, cholestasis, or tumor growth. Although liver transplant is a lifesaving event, challenges for these patients do not end with a completed operation. The complex vascular reconstruction required in transplanting a new liver graft results in high risks for organ loss due to thrombosis, occlusion, and/or stenosis in the reconstructed vessels.

Despite improved procedural techniques and careful observation, vascular complication rates remain high and cause measurable patient harm. Early hepatic artery thrombosis occurs in ~5% of transplant recipients and is implicated in 53% of graft losses and 33% of deaths in the early postoperative period. Late hepatic artery thrombosis is associated with transplanted organ failure and sepsis. Late hepatic artery thrombosis also has a devastating effect on the lining of the bile ducts leading to liver graft loss. Similarly, hepatic artery stenosis occurs often within 3 months in 11% of transplantation recipients and can develop into biliary ischemia, which again may lead to liver graft loss.[1]

Given these risks, patients are screened regularly post-transplantation using spectral Doppler ultrasound (US) imaging. To aid in diagnosis, several quantitative measures have been proposed to quantify risk, including the resistive index (RI), stenosis index (SI; ratio of area under the high-frequency signal to low-frequency signal in the spectral Doppler), acceleration time (AT), and peak systolic velocity.[2] However, as false-negative USs can result in disastrous consequences, including liver graft loss, a conservative threshold is often taken, resulting in false-positive rates as high as 27% to 40%. For example, RI (the most commonly used single metric) has a shown consistent sensitivity of 60% to 62% and specificity of 77% to 80% for detecting arterial stenosis across a body of literature.[3–7] Patients with a positive US screening must then undergo an invasive angiogram or computed tomography angiogram. However, the angiogram itself presents several considerations, first, a monetary cost ranging from ~ $15,000 to $30,000, and second, a 5% to 10% major complication rate including bleeding, clotting, vessel injury, and infection.

As such, there exists a significant incentive to improve screening specificity and reduce unnecessary angiograms without comprising the screening test's sensitivity in identifying patients who need intervention. We posit such a shift is possible through the utilization of data extracted from the complete Doppler waveform. Doing so would greatly expand the quantity

Received for publication January 20, 2022; accepted July 20, 2022.

*Division of Health Services and Outcomes Research, Children's Mercy Kansas City, Missouri; †Department of Pediatrics, University of Missouri-Kansas City School of Medicine; ‡Department of Biomedical Engineering, Northwestern University, Chicago, IL; §Department of Medicine (Biomedical Informatics), Stanford University, Palo Alto, CA; ‖Department of Radiology, Children's Mercy Kansas City, Missouri; and ¶Department of Radiology, University of Missouri-Kansas City School of Medicine, Missouri.

The authors declare no conflict of interest.

Address correspondence to: Sherwin S. Chan, MD, PhD, Department of Radiology Pediatric Radiologist, Children's Mercy Hospital, 2401 Gillham Rd, Kansas City, MO 64108 (e-mail: sschan@cmh.edu).

DOI: 10.1097/RUQ.0000000000000624

of data that must be analyzed at the point of diagnosis, introducing challenges for any individual to assess multiple sources of potentially differing results. Data-driven machine learning techniques are well suited for such a task and have become increasingly common in the radiology field. Often associated with deep learning, extensive work has demonstrated success for image analysis of various liver conditions.[8,9] However, recent works on more interpretable machine learning approaches have been applied in a variety of hepatology tasks ranging from ability to predict graft failure at the time of transplant,[10] to early detection of patients with non-alcoholic fatty liver disease, to prediction of acute kidney injury post-transplant or cirrhosis outcomes for viral hepatitis.[11] Using an extraction technique developed by our group, this work presents a novel interpretable approach to identifying stenosis and occlusion in post-transplantation patients.

In line with the Checklist for Artificial Intelligence in Medical Imaging,[12] This article begins with a comprehensive description of the data, feature engineering, and study design. Next, it outlines a novel framework for the prediction of stenosis and occlusion using data derived from the complete waveform. From there, detail is provided around the evaluation of the proposed framework, for both performance and interpretability. This article concludes with a discussion of the clinical implications of the framework results and highlights ongoing work to push techniques closer to practice.

MATERIALS AND METHODS

Study Design and Data

This article undertakes a secondary retrospective analysis of an existing deidentified dataset of Doppler US waveform data collected from January 1, 2006, to December 31, 2010, from a single large tertiary medical center. The data were collected by first identifying all patients who underwent mesenteric catheter angiography using a searchable index of radiology reports (zVision, Clario Medical, Seattle, WA). Reports were then reviewed by members of the study team to identify any patient with hepatic allografts who underwent angiographic evaluation of the transplanted hepatic arteries. All subject data were collected retrospectively, under a University of Washington–approved institutional review board protocol with a waiver of informed consent and deidentified upon release from the study center. The study was approved by a local institutional review board, and requirement for informed consent was waived.

From these, every patient who underwent spectral Doppler US of the transplanted hepatic arteries within 30 days preceding their angiographic procedure was included. However, these patients alone may represent an inherently biased sampled, as those who underwent angiography may have done so as they were considered high probability to have stenosis by US. Also, patients whose arteries were found to be occluded on angiography may have had a patent arterial system on the preceding US or had collateral circulation. To better study US-driven stenosis identification, we also performed extensive chart review to identify a population that was considered low risk to have stenosis by US and had anatomic imaging to verify they truly did not have stenosis. This search identified a cohort of patients who had liver transplants and contrast-enhanced computed tomography (CT) and Doppler transplant US evaluation in the 30 days before the CT examination. We chose 60 consecutive patients in this cohort for inclusion. In this cohort, each CT was evaluated by a board-certified radiologist to ensure no hepatic arterial abnormalities were present (stenosis or occlusion).

In total, the search produced a cohort of 159 liver transplant recipients. All transplants were full orthotopic liver transplants. A set of exclusion criteria were then applied. Specifically, we excluded patients who did not have sufficient quality spectral Doppler tracings from both the left and right hepatic arteries. We defined sufficient quality waveforms as ones that contained 3 consecutive accurate waveforms without signal loss. We chose this metric because this waveform quality is required to accurately calculate one of the baseline measures (SI). Substantial noise included breathing artifact that causes signal dropout and movement of the artery during acquisition; an example of an invalid waveform can be found in Figure 1.

After exclusion, a final cohort of 129 patients remained (75 without pathology, 54 with stenosis/occlusion) for analysis. In addition to the waveform, patients' age and sex were recorded. An overview of demographics for the final cohort by outcome can be found in Table 1.

Data Preprocessing and Feature Engineering

We used the available spectral Doppler waveform recordings (~7 seconds) in hepatic transplant screening USs and applied a feature engineering process to derive meaningful measures of the waveform shape. We began with the derivation of common clinical metrics such as AT, SI, and RI, as well as left and right half-life to capture the rate of rise of the systolic wave and the rate of decline of the diastolic wave (defined in Table 2). Next, utilizing work previously published by our group, the waveform was decomposed, from which the top 10 principal components were extracted to capture quantitative measures of waveform shape.[13] Together, these features will provide model-based approaches a more comprehensive view of an individual patient as compared with any single screening metric. A complete listing of variables and definitions can be found in Table 2.

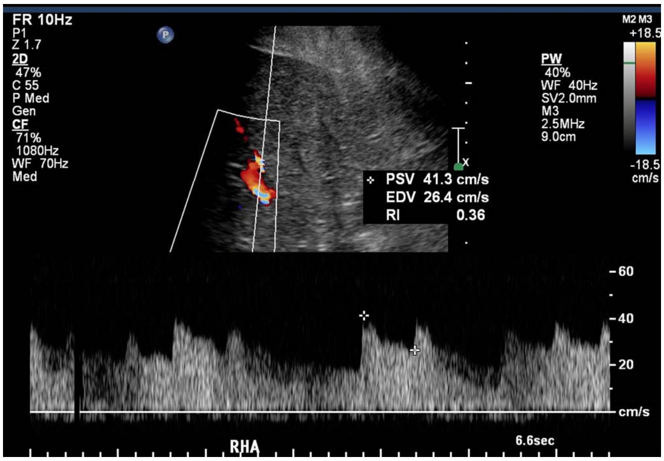


FIGURE 1. Spectral Doppler waveform of the right hepatic artery in a 63-year-old man with right hepatic artery stenosis by angiography. This is an example of excluded (invalid) waveform. Note the dropout of the spectral Doppler signal as the patient breathes.

TABLE 1. Demographic Attributes of Study Cohort

	No Stenosis/Occlusion (n = 79)	Stenosis/Occlusion (n = 54)	*P**
Age, mean (SD)	55.59 (9.41)	53.52 (9.20)	.025
Sex	Male: 82.67%, female: 17.33%	Male: 68.52%, female: 31.48%	.090

*Statistical comparisons were made using Mann-Whitney *U* and Fisher's exact tests for age and sex, respectively.

It is common for multiple waveforms to be recorded in a single screening. Multiple samples allow increased confidence in evaluating the waveform shape. To guard against the potential bias of sonographers recording more images in challenging or worrisome cases, the values derived across multiple waveforms in collected in an imaging series were collapsed into a single vector per patient. Rather than simply take a mean value, aggregation operations were selected to align with relationships between each feature and the clinical pathology of restricted blood flow that arises due to stenosis or occlusion. As US is a screening examination, we use the most worrisome value to summarize risk. Specifically, the minimum value of each metric was taken for the SI and RI to capture reduced measures of flow, as well as all derived principal component features. The maximum values of the left and right half-lives and AT were used as those metrics are expected to be higher in cases of stenosis or occlusion. Finally, the mean frequency was used as it corresponds to a patient's average heart rate.

Ground Truth and Outcome Definition

Angiography images were rigorously evaluated to determine ground truth for stenosis and/or occlusion. An interventional radiologist examined each angiogram for pathology and compared their impression with the clinical report. If there was a discrepancy between the report and review, a second reader was used to break the tie. For each subject, positive outcomes were assigned for significant angiographic stenosis (>50% stenosis) requiring intervention, angioplasty, and/or stenting, as well as hepatic artery occlusion. In addition to reviewing imaging studies for patients who underwent only CT, the subject's medical records were also reviewed until December 31, 2010, to ensure they did not undergo later intervention.

METHODS

Taking the vector of derived waveform features for each patient, the proposed framework is centered on the specification of an Extremely Randomized Tree (ERT) model.[14] Similar to random forests, ERTs extend the concept of a single decision tree into an ensemble paradigm, where many trees are trained on various combinations of features and instances. However, in contrast with random forests, ERTs are designed to split at a random threshold within each feature, rather than utilizing a best splitting criterion to optimize tree-building. As our dataset contains known predictors of stenosis in the SI and RI features, which would likely dominate Gini/Entropy-based measures of feature importance, random splitting offers a means to reduce the variance of the final model and further improve generalizability for the proposed framework.

Extremely Randomized Tree models are composed of several parameters known to have a significant impact on the overall model performance.[15,16] As such, the methods below present a framework for tuning the model to the improve false-positives rates while maintaining high levels of sensitivity in identifying cases of stenosis and occlusion. This approach is broken into 2 primary elements: first, testing a range of hyperparameters used to define the ERT structure (eg, number of trees in the forest, how many waveform features to consider in each tree), and second, within each hyperparameter set, determination of the optimal threshold for the estimated model probability to define a stenosis/occlusion case. This allows us to set an acceptable sensitivity level and utilize a bootstrap approach across training data to identify a threshold of stenosis/occlusion probability that maximize specificity. Detailed methodology of each component is provided in the sections to follow, whereas a visual overview of the framework can be found in Figure 2.

Model Specification and Hyperparameter Tuning

First, a grid search was performed encompassing the following features [possible values]: the number of trees [1000, 2000], number of features used [4, 8, 12], maximum depth of the tree [3, 5, unbounded], minimum number of samples required to split a node [5, 10, 15, 30], and the minimum number of samples required to be at a leaf [5, 10, 15, 30]. In total, 288 distinct parameter combinations were evaluated in parallel.

Classifier Threshold and Grid Search Evaluation

Next, for each configuration in the grid search, model performance was compared through measures of sensitivity (true-positive rate) and specificity (true-negative rate). However, given the screening nature of this evaluation, and the significant adverse cost of false-negatives, it is likely the default classifier threshold (0.5) is insufficient to meet the current clinical state-of-the-art. Adjustment of this threshold is often done post hoc by identifying a satisfactory balance of sensitivity/specificity using a receiver operating curve (ROC). However, as we are evaluating across hundreds of dependent datasets (same data, different model parameters), repeatedly assessing performance can represent a form of data snooping. Rather, an internal validation approach was used to do so in less biased manner.

First, within each hyperparameter configuration, a 1000-iteration bootstrap was conducted. Bootstrapping is a statistical technique in which instances are sampled with replacement until the size of the original dataset is reached. In doing so, approximately 37% of instances are known to be excluded, known as out-of-bag (OOB) samples.[17] By training an ERT model on the resampled data, we were able to compute the ROC using the OOB samples as an ad hoc test set. Then, starting at the highest sensitivity, iterate backward to determine the lowest threshold at which sensitivity exceeded the clinical state-of-the-art performance. Here, state-of-the-art sensitivity was considered to be 94%, in line with the performance of SI as found by Le et al.[2] The threshold value of the ROC was

TABLE 2. Complete Set of Features and Definitions Used in the Study

Variable Name	Variable Description
Age	Age of patient at the time of ultrasound
Sex	Sex of patient
Resistive index	(Peak systolic velocity – end diastolic velocity) / peak systolic velocity
Stenosis index	Ratio of sum of amplitudes of the high-frequency components of waveform / amplitude of the fundamental frequency of the waveform
Acceleration time	Time between end diastolic velocity and early systolic peak during the systolic upstroke
Right half-life	The half-life is defined as the time it takes for the velocity of the blood in the measured artery to travel half the amplitude from the high point to the low point.
Left half-life	The half-life is defined as the time it takes for the velocity of the blood in the measured artery to travel half the amplitude from the low point to the high point.
Frequency	Frequency of the waveform, which is also the average heart rate of the patient during waveform acquisition
PCA (1 to 10)	PCA results from axes 1 to 10

PCA, principal component analysis.

recorded, and the process was repeated. At the conclusion of the bootstrap iterations, the median threshold was selected.

As the threshold itself is now a parameter estimate, it was important to compute the expected sensitivity and specificity to ensure that the state-of-the-art sensitivity can still be achieved and to measure the expected specificity gains through the model-based approach. As such, the 1000-iteration bootstrap was repeated using the same randomization; however, rather than use OOB samples to explore the range of predicted probabilities, the predictive performance of the OOB samples at the identified threshold was evaluated as a binary classification.

Once the grid search was completed, the optimal hyperparameter set was selected as the configuration with the highest sensitivity. In the event that 2 or more configurations had identical specificity performance, that with the highest estimated sensitivity was selected. Should there remain ties across both metrics, a configuration would be selected at random. The ERT with the selected hyperparameters was then retrained on the complete training data and evaluated as detailed below. All analyses were completed using Python 3.7.6, Pandas 1.0.1,[18] SciPy 1.4.1, NumPy 1.18.1,[19] and scikit-learn 0.22.1.[20]

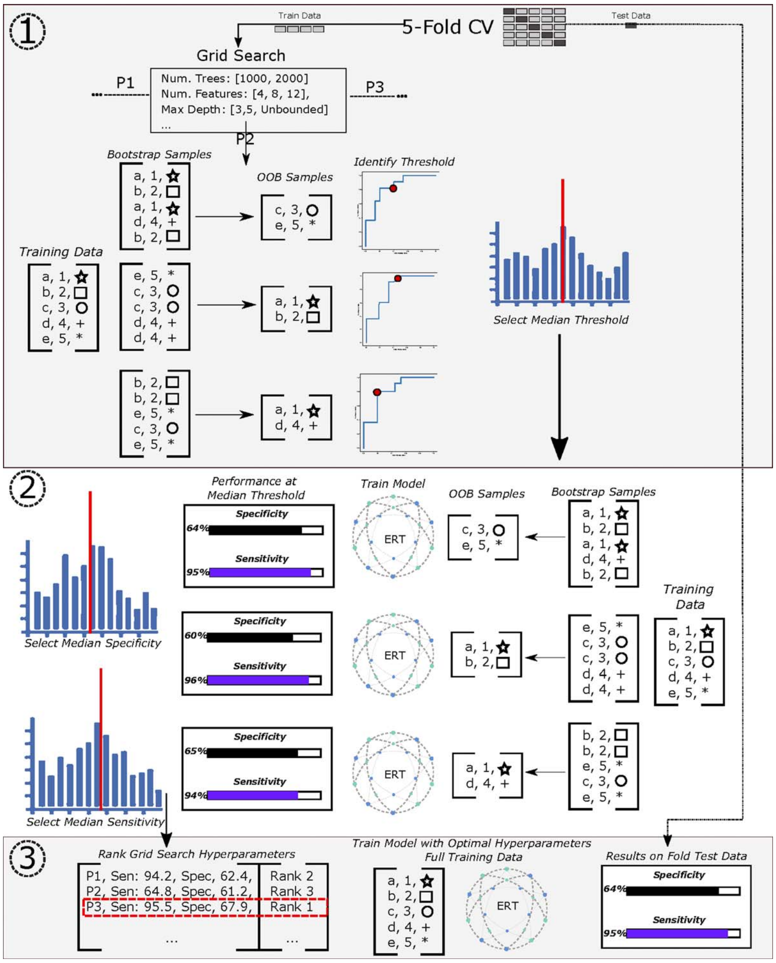
Evaluation

To assess the generalized performance of the proposed methodology, 5-fold cross-validation (CV) was used. Given the natural imbalance between outcome classes, stratified K-fold was utilized to ensure each test set contained a representative proportion of stenosis and occlusion cases as was seen across the study cohort. To prevent any data leakage, test data in each fold were never used in any aspect of the framework until evaluation on the final parameter configuration, and an independent grid search and bootstrap evaluation were repeated across the subset of training data of each fold. In line with recent literature to improve the robustness of CV results, 10 independent iterations of 5-fold CV were performed, each with a different random seed and subsequently test/train splits. Such work has empirically demonstrated summarization of performance *across* 10 independent runs, averaging within the 5 folds of each independent CV, and provides a more reliable estimate of model generalization than assessing the variance across the 5-fold CV or the performance across a single 10-fold CV alone.[21] On each of the same testing datasets, we also implemented a series of model baselines using measures of RI and SI. These included evaluating commonly clinically accepted thresholds of RI < 0.5, RI < 0.4, and SI < 1.35. In addition, for the RI metric, we replicated the model approach of utilizing the training data to identify a threshold with a sensitivity of 94% and evaluated the distribution of specificity values across the repeated K-folds.

Model Interpretability

Although aiding in generalizability, the ensemble nature of the forest of discrete trees produced by the ERT model precludes outputting of a single decision pathway for any one test subject. However, providing insight into the model's operation is paramount for clinical acceptance of any such tool. To do so, we took 2 approaches to elucidate model behavior.

First, at a macrolevel, we computed feature importance in the ERT model selected from the grid search and fit across the full set of training data, before testing. Specifically, we utilized the Gini



1
Train Data
5-Fold CV
Test Data
Grid Search
P1
Num. Trees: [1000, 2000]
Num. Features: [4, 8, 12],
Max Depth: [3,5, Unbounded]
...
P3
P2
Bootstrap Samples
OOB Samples
Identify Threshold
Training Data
Select Median Threshold
2
Performance at Median Threshold
Train Model
OOB Samples
Bootstrap Samples
Specificity
Sensitivity
64%
95%
60%
96%
65%
94%
ERT
Training Data
Select Median Specificity
Select Median Sensitivity
3
Rank Grid Search Hyperparameters
P1, Sen: 94.2, Spec, 62.4, Rank 2
P2, Sen: 64.8, Spec, 61.2, Rank 3
P3, Sen: 95.5, Spec, 67.9, Rank 1
Train Model with Optimal Hyperparameters Full Training Data
Results on Fold Test Data

FIGURE 2. Workflow of the proposed framework. 1, Training data from a CV fold is run through a grid search. For each parameter, configuration bootstrap sampling is used on the CV-fold's training data to determine a threshold at which state-of-the-art sensitivity is obtained using out-of-bag samples. 2, For the same grid search parameters, this threshold is used to estimate expected specificity of the model. 3, Grid search results are sorted and the configuration maximizing specificity (ties broken by highest sensitivity) is then used to train a final model run once to quantify performance on held out test data for the respective fold. The process is then repeated for each of the 5 CV test folds, and then overall across the 10 independent CV runs to estimate overall performance.

impurity importance and recorded the distribution of feature scores across each of the 50 runs.

Second, at an individualized level, we utilized the model interpretability package SHAP to present a case study demonstrating how Shapley values can provide insight into the specific features that drive an outcome prediction for an unseen test instance.[22] Case study patients were selected from a random test set and included a stenosis case, a control case, and a misprediction. Utilizing the final ERT model, a SHAP explainer was trained to learn the association between model output and the magnitude/directionality of feature values. Each test case was explored individually.

RESULTS

Given the intended use case of the Doppler US as a hepatic screening test for posttransplantation patients, the results of the proposed framework summarize the performance of both sensitivity and specificity[2] in the detection of stenosis and/or occlusion. Utilizing the aggregation recommendations, the results of both metrics were collapsed *within* the runs comprising a single 5-fold validation. Comprehensive summary statistics were then computed for each metric across each of the 10 independent runs, including measures of central tendency (mean), measures of variability (standard deviation) as well as the minimum and maximum for a sense of overall performance distribution.

The results of this process can be found in Table 3. When compared with the threshold of RI < 0.5 in this dataset, the proposed model was found to achieve significantly higher sensitivity (0.94 vs 0.85) compared with specificity (0.65 vs 0.75). In addition, when RI and the model were tested using similar conditions, the sensitivities of both tests were similar but the model had higher specificity compared with RI alone (65% vs 30%).

For the readers' convenience, we also present boxplots of the performance across the 5 folds of each of the independent

TABLE 3. Performance Results of the Machine Learning Model

		Sensitivity	Specificity
Model	Mean (SD)	0.95 (0.01)	0.65 (0.03)
	Min/Max	0.93/0.96	0.60/0.68
RI constrained	Mean (SD)	0.94 (0.01)	0.30 (0.03)
	Min/Max	0.93/0.95	0.26/0.35
Discrete threshold	RI < 0.5	0.85	0.75
	RI < 0.4	0.63	0.84
	SI < 1.35	0.87	0.52

These values represent the distribution of results across the 10 independent runs (j) of the framework. Each run encompasses a 5-fold CV (k), from which the mean value is used as per the j-k-fold evaluation technique for improved estimates of generalizability. Baseline RI and SI are provided utilizing identical j-k-fold test. RI constrained represents average RI performance when constrained to achieve the same 94% sensitivity as the model. SD/Min/Max are not provided across the folds as results were stable (SD < 0.01) with Min/Max within 0.01.

Max, maximum; Min, minimum.

CV runs. Notably, the results are highly stable, suggesting reliability in the threshold selection training process (Fig. 3).

Model Interpretability

Macro-level

We first looked at global measures of ERT feature importance. The mean importance for each feature was computed across the 5 folds of each run. We then examined the distribution of these importances across the 10 runs using the boxplots found in Figure 4.

Individual Level

Looking next to an individual subject, SHAP allowed us to shed light on how the waveform attributes of a patient contribute to the probability of stenosis as estimated by the ensemble.

To do so, we utilized representations known as force plots.[23] These figures quantify the culminative effect of each feature's contribution to the final prediction, using a type of tug-of-war approach. Utilizing the relationships learned between feature values and outcomes during model training, arrows pointing right indicate the value of a patient's respective feature would increase their probability of a stenosis/occlusion outcome, whereas arrows pointing left indicate increased probabilities of a negative outcome. The length of the arrow is relative to the overall magnitude of the feature's contribution to the outcome determination as defined by its SHAP value. Results of this process can be found in Figure 5A and B for a positive and negative example, respectively. An example of a misprediction can be seen in Figure 5C, where the model predicted a stenosis diagnosis, with ground truth of a negative patient.

DISCUSSION

This study has presented an automated machine learning framework for the detection of clinically significant stenosis and occlusion in hepatic arteries after liver transplantation. A more comprehensive set of features derived from the full spectral Doppler US waveform was used as compared with common quantitative point estimates used today. Our approach was able to achieve a sensitivity and specificity of 94% and 65%, respectively. This is an improvement over current SI and RI metrics alone, which have sensitivities of 94% to 96% and 60% to 62% and specificities of 29% to 52% and 77% to 80%, respectively. The specificity of the machine learning algorithm is objectively low, but this dataset is also biased toward patients who were extremely difficult to diagnose by US. This is evidenced by the fact that 58% (75/129) of the patients went on to angiography after their US examinations.

Notably, the clinical implications of increasing specificity while maintaining state-of-the-art sensitivity are considerable, namely a decrease in false-positives that currently go on for angiographic evaluation. This in turn would save patients from undergoing unnecessary procedures and save money. The ideal implementation of this model would be achieved through collaboration with US manufacturers to include this measure on their machines. The second possible implementation would be

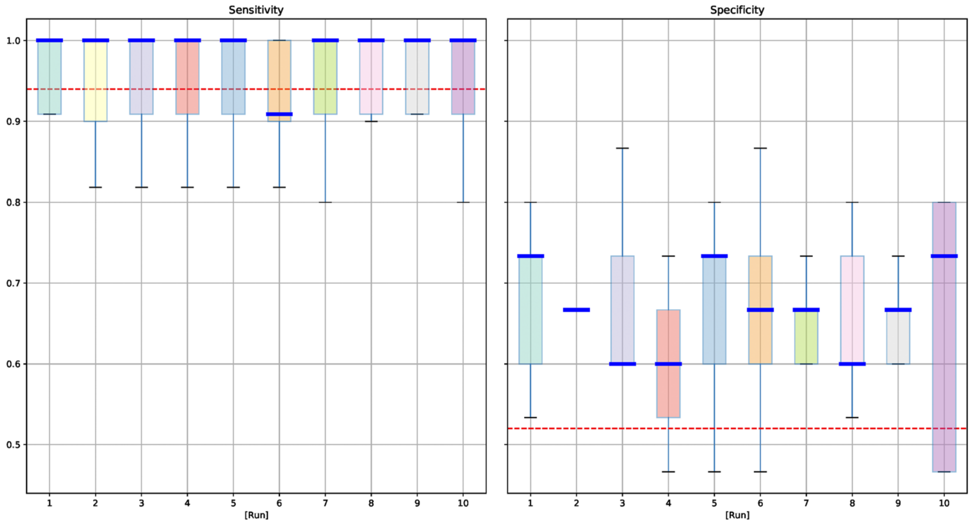


FIGURE 3. Distributions of sensitivity and specificity across the 10 independent runs of 5-fold CV. Dark blue lines indicate the median value. Red dashed lines represent the SI performance: sensitivity, 0.94; specificity, 0.52.

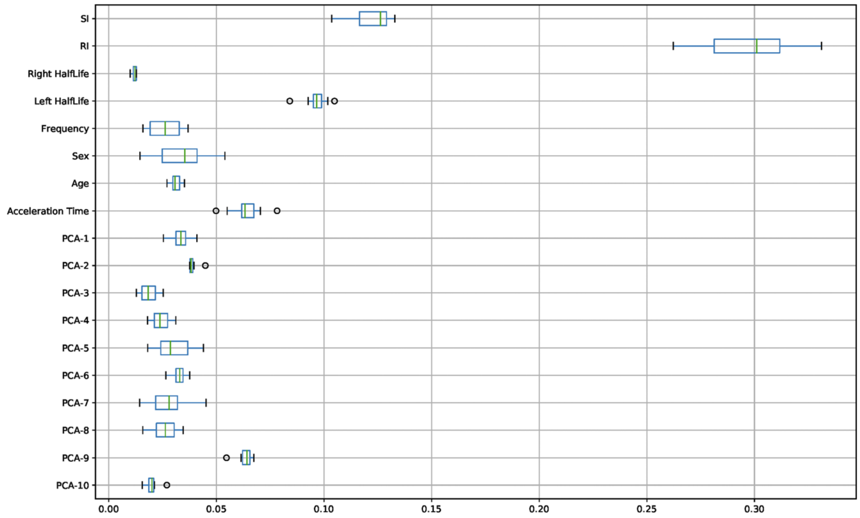


FIGURE 4. Distribution of mean feature importance across the 5-folds of each of the 10 CV runs. PCA (1 to 10) indicates principle component analysis axis.

to create stand-alone software to calculate the risk of pathology from the already acquired waveforms. In either case, the implemented software could calculate a prediction from the waveform in seconds, making this easy to integrate into normal clinical practice.

Moreover, a primary advantage of our proposed framework over deep learning models is the potential for explainability rather than simply a probability estimate. Using the approximation techniques of SHAP, the case study on patient-level interpretability highlights several interesting findings. For example, in the correct stenosis prediction (Fig. 5A), the patient's age and SI actually decrease the stenosis probability (as indicated by the left-facing arrows). However, the magnitude of the RI, together with aspects of half-life and AT, was strongly related (noted by the length of the array) to the positive class and the model's overall correct positive prediction. Similarly, in the misprediction, we find that RI and left half-life were somewhat indicative of a negative screening (small, left-facing arrows), whereas SI was strongly associated with a positive outcome based on the training data ERT. Thus, by offering tangible insight into specific features driving risk for an individual patient, it is possible to help guide care plans and practitioner focus, allowing a radiologist to easily identify areas for further evaluation or offer a compelling reason to justify disagreement with the algorithm.

Limitations

It is important to note 2 limitations regarding the features and cohort. First, the use of principal component analysis as a part of the model does impede the interpretability of the model when those components are a large part of the individual prediction. This is because principal components do not have a clear waveform correlate. In the future, we intend to explore the use of alternative dimensionality reduction approaches such as factor analysis to improve interpretability of the model.

Next, although an effort was made to collect negative samples from both those who underwent angiography and those not referred for additional screening, the use of only 60 CT controls may have influenced the imbalance ratio. In a screening population, the controls will greatly outnumber patients with stenosis/occlusion. Therefore, the study population has a higher rate of positives relative to the normal screening population. This is not a major concern, as the model can easily be applied for those whom traditional screening indicates the need for invasive testing, before angiography is performed; we are actively collecting data across the complete set of US screening examinations for another regional medical center to validate these results on a larger population.

Finally, it is important to note that as with other US techniques, performance of the proposed method is reliant on the ability to capture reliable waveforms. This is definitely a limitation as the transplant hepatic artery is difficult to image. However, if this method becomes more proven and if it is used clinically, sonographers will likely spend more time getting waveforms that can be input into the model. An area of future work is also creating an algorithm to grade the reliability of waveforms before consumption by the model.

CONCLUSION AND FUTURE DIRECTION

Though this work, we have introduced an automated machine learning framework that improves the specificity for predicting the presence of hepatic stenosis or occlusion utilizing a spectral Doppler US waveform while maintaining a high sensitivity. Implementing this prediction algorithm could reduce the need for potentially unnecessary invasive procedures, while accounting for the extremely high cost of false-negatives in the potential organ loss.

However, this is only the first step in achieving clinical utilization of this technology. Larger samples from national transplant populations are required to rigorously determine a singular set of model parameters that will be required for implementation of pretrained models into existing software.

We are concurrently exploring 2 approaches for extending the methodology. First, we are working to integrate data captured

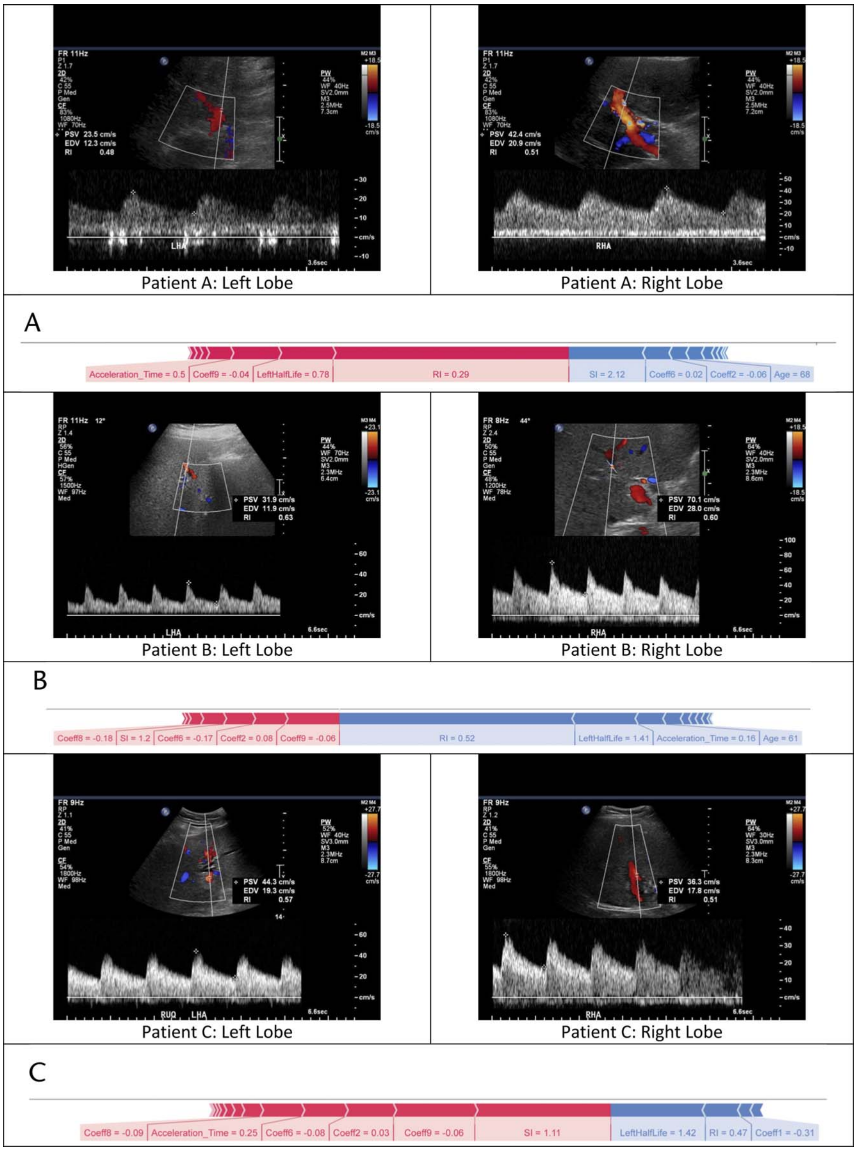


FIGURE 5. A, Spectral Doppler waveforms of the left and right hepatic artery in a 68-year-old man with proper hepatic artery occlusion on angiography. A force-plot highlighting a correct positive prediction of stenosis. Results were strongly influenced by the values of the RI and left half-life with respect to the expected distributions of training data for stenosis-positive patients, whereas values of the SI and some Fourier transformed PCA data provide weaker indications of a negative reading. B, Spectral Doppler waveforms of the left and right hepatic artery in a 61-year-old man with no hepatic artery occlusion or stenosis by follow-up CT. A force-plot highlighting a correct negative prediction of no stenosis. Results were strongly influenced by the values of the RI, left half-life, and AT within the expected distributions of negative/control patient's training data, whereas some values of the Fourier transformed PCA data provide weaker indications of a stenosis reading. C, Spectral Doppler waveforms of the left and right hepatic artery in a 56-year-old man with no hepatic artery occlusion or stenosis by follow-up CT. A force-plot a misprediction of stenosis for a negative patient. Although the RI and left-half-life values for this patient were supportive of a negative prediction, together, values of the SI and several dimensions of the Fourier transformed PCA data outweighed the final result for a model prediction for a positive outcome. In all cases of force plots, the direction (color) and size of each arrow are based on the relative contribution of the feature value. Red arrows pointing right indicate that, in reference to the model training data, the feature value increases the probability of a stenosis prediction, whereas blue arrows pointing left increase the probability of a negative prediction. The size of the arrow indicates a measure of magnitude for the contribution.

from discrete imaging waveforms, such as studies with differing probe positions, rather than aggregating data into a single instance. This approach can potentially allow for fine-grained interpretability by directing the reading radiologist to review specific imaging results. Second, we are looking to develop a longitudinal component to the ensemble, where changes in parameter values *between* screenings can be used to better quantify early signs of stenosis or occlusion for an individual.

REFERENCES

1. Caiado AH, Blasbalg R, Marcelino AS, et al. Complications of liver transplantation: multimodality imaging approach. *Radiographics*. 2007;27(5):1401–1417.
2. Le TX, Hippe DS, McNeeley MF, et al. The sonographic stenosis index: a new specific quantitative measure of transplant hepatic arterial stenosis. *J Ultrasound Med*. 2017;36(4):809–819.
3. Dodd GD, 3rd, Memel DS, Zajko AB, et al. Hepatic artery stenosis and thrombosis in transplant recipients: Doppler diagnosis with resistive index and systolic acceleration time. *Radiology.* 1994;192(3):657–661.
4. Platt JF, Yutzy GG, Bude RO, et al. Use of Doppler sonography for revealing hepatic artery stenosis in liver transplant recipients. *AJR Am J Roentgenol*. 1997;168(2):473–476.
5. Sidhu PS, Ellis SM, Karani JB, et al. Hepatic artery stenosis following liver transplantation: significance of the tardus parvus waveform and the role of microbubble contrast media in the detection of a focal stenosis. *Clin Radiol*. 2002;57(9):789–799.
6. Tamsel S, Demirpolat G, Killi R, et al. Vascular complications after liver transplantation: evaluation with Doppler US. *Abdom Imaging*. 2007;32(3):339–347.
7. Vit A, De Candia A, Como G, et al. Doppler evaluation of arterial complications of adult orthotopic liver transplantation. *J Clin Ultrasound*. 2003;31(7):339–345.
8. Kalyan K, Jakhia B, Lele RD, et al. Artificial neural network application in the diagnosis of disease conditions with liver ultrasound images. *Adv Bioinformatics*. 2014;2014:708279.
9. Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. *World J Gastroenterol*. 2019;25(6):672–682.
10. Lau L, Kankanige Y, Rubinstein B, et al. Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation*. 2017;101(4):e125–e132.
11. Spann A, Yasodhara A, Kang J, et al. Applying machine learning in liver disease and transplantation: a comprehensive review. *Hepatology.* 2020;71(3):1093–1105.
12. Mongan J, Moy L, Kahn CE, Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029.
13. Baraboo JJ, Dinakarpandian D, Chan SS. Automated prediction of hepatic arterial stenosis. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:58–65.
14. Geurts P, Ernst D, Wehenkel L. Extremely Randomized Trees. *Mach Learn*. 2006;63(1):3–42.
15. Probst P, Boulesteix A-L, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. 2018. arXiv:1802.09596. https://ui.adsabs.harvard.edu/abs/2018arXiv180209596P. Accessed February 1, 2018.
16. Olson RS, Cava WL, Mustahsan Z, et al. Data-driven advice for applying machine learning to bioinformatics problems. *Biocomputing*. 2018;23:192–203.
17. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc*. 1983;78(382):316–331.
18. McKinney W. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference. 2010;445(1):51–56.
19. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–272.
20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. 2012. arXiv:1201.0490. https://ui.adsabs.harvard.edu/abs/2012arXiv1201.0490P. Accessed January 1, 2012.
21. Moss H, Leslie D, Rayson P. *Using J-K-fold Cross Validation to Reduce Variance When Tuning NLP Models*. In Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico: Association for Computational Linguistics; 2018:2978–2989.
22. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Luxburg Uvon, Guyon I, eds. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc; 2017:4768–4777.
23. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749–760.