ORIGINAL ARTICLE

Predicting chatter using machine learning and acoustic signals from low-cost microphones

Sam St. John[1] · Matthew Alberts[1] · Jaydeep Karandikar[2] · Jamie Coble[3] · Bradley Jared[4] · Tony Schmitz[2,4] · Christoph Ramsauer[5] · David Leitner[5] · Anahita Khojandi[1]

Abstract

Machining chatter is a phenomenon resulting from self-oscillation between a machining tool and workpiece. This self-oscillation results in variation on the machined product that reduces the ability to meet desired specifications. Chatter is a widely studied topic as it directly relates to the quality of machined products. This study details the application of a Random Forest (RF) classifier with Recursive Feature Elimination (RFE) to machining audio collected by a single microphone during down-milling operations. This approach allows straightforward feature elimination that results in an easily understood set of analyzed dimensions. Stability is predicted solely based on the classification output of the RF classifier. Our approach proves highly predictive with consistent machining setup and a small sample set. We also review transferability between machining setups and present key findings. Our RF approach demonstrates the ability to analyze and classify chatter through a low-cost approach with limited training data required. The motivation for using a single microphone is to enable detection on machines without other sensors, such as accelerometers, present in the machining setup. The value of the in-process sensor and chatter classifier is highlighted because the machining setup included asymmetric dynamics that reduced the accuracy of the traditional analytical stability solution. We see a natural progression to deploying this audio-only methodology with real-time processing and classification using either a laptop or smartphone. This progression will allow visual indicators during the machining process that can alert machinists of progression into unstable machining processes.

Keywords Machine learning · Random forest · Recursive feature elimination · Chatter · Stability · Advanced manufacturing · Low-cost sensors

1 Introduction

Chatter identification and remediation is a long-standing research topic [1, 2] that has resulted in common approaches for detection of chatter. A common approach for online

✉ Anahita Khojandi
  khojandi@utk.edu

Extended author information available on the last page of the article.

detection of chatter uses a Fast Fourier Transform (FFT) of an audio signal. The FFT results allow comparison with the calculated Tooth Passing Frequency (TPF). Significant peaks not aligned with the TPF or its harmonics (multiples) indicate chatter [3]. A common analytic approach is the creation of stability lobe diagrams. These diagrams illustrate the relationship between spindle speed, typically on the horizontal axis, and depth of cut, typically on the vertical axis. Stability lobe diagrams reveal an area of expected chatter above the stability boundary and an area of expected stability below the boundary [2–6]. Figure 1 provides an example of a stability lobe diagram. Machinists can use stability lobe diagrams to select parameters that avoid chatter given the machining system's dynamic response and the cutting force model. The traditional stability lobe diagram is deterministic and does not include uncertainties in the diagram inputs, however, stability lobe diagrams are not definite [7]. These diagrams include a region of uncertainty in the transition between stable and unstable

Fig. 1 Example stability lobe diagram

areas and may include lenses and islands of chatter within stable areas [5, 6]. Researchers have attempted to address this through the use of a reliability zone that can confidently describe where chatter will be absent [8]. Even with this advance, machining configuration based on stability lobe diagrams and reliability zones requires calculation in advance of machining with detailed knowledge of the tool setup.

Chatter detection is expanding beyond peak analysis and stability lobe diagrams. Significant research is available regarding the use of sensors and associated calculations to detect chatter either during or after the machining process. In the review article [9], the authors classify sensor-based methods based on data acquisition and signal processing methods. Data acquisition methods include the use of acceleration, force, sound, current, and image signals. Processing methods include time domain, frequency domain, and time-frequency domain methods.

The ability to collect rich signal information during the machining process is driving further research leveraging machine learning and deep learning techniques. The signals used align with data acquisition methods previously described. Key examples of data acquisition for chatter detection include the use of accelerometers [10], dynamometers [11, 12], acoustic emission sensors [13–15], microphones [16], drive current derived from CNC commands [17], and online image capture [18]. Although all data acquisition methods provide benefits, the use of microphones is comparably inexpensive and unobtrusive. Key advantages for microphone capture are sensitivity to chatter in low-force situations and ability to isolate the sensor from machining structure. Limitations include microphone bandwidth and environmental noise [19].

Data-driven methods for these signal types represent a broad research topic. Researchers have applied Support Vector Machine (SVM) [10], Extra Trees Classifier (ETC) [20], and Deep Multi-Layer Perceptron (DMLP) [21] approaches to accelerometer data for detecting chatter. A Perceptron Artificial Neural Network (ANN) classifies chatter with acoustic emission data [22]. Researchers have applied statistical variance analysis [19], as well as Hidden Markov Model (HMM) and SVM [23] approaches with microphone data. K-means clustering [24] and Long Short-Term Memory (LSTM) [25] identify milling conditions using drive current signals [24].

Microphone data has been combined with image processing by displaying and interpreting microphone data with spectrogram images. An example of this is [16], in which the authors compare a back propagation neural network (BP-NN) and convolutional neural network (CNN) based on the ability to classify chatter. The BP-NN is based on traditional sound signals, including amplitude, frequency and power coefficients. The CNN analysis is based on a spectrogram image that is a visual interpretation of the sound signal including details regarding frequency, time domain and magnitude of signal at frequency and time domain combinations. Popular CNN projects for handwriting recognition and image processing use relatively small image sizes, such as 32x32 pixels. This results in loss of detail in the features for chatter detection. To avoid this loss, the authors use image sizes of 150x150 pixels. Both BP-NN and CNN classification results show high predictability with accuracy above 90%. Overall, the CNN model performs better than the BP-NN model.

In [26], the authors provide another example of the convergence of microphone signal data and image processing techniques. Similar to the prior CNN work, this study uses a spectrogram derived from audio signals. This approach differs from the previously described work in the application of an autoencoder to remove noise from the data and reduce dimensionality. The researchers show an ability to reduce the number of required predictive dimensions to two derived dimensions. Once the autoencoder reduces dimensionality, the researchers use a SVM to classify the presence of chatter. The resulting data shows a delineation between audio files with chatter and without. This study indicates that the image data derived from audio signals is highly compressible and this compressed data is viable for the classification of chatter.

Although the prior data-driven approaches to chatter identification represent important advances, they also have limitations. Each approach represents a "black box" method with limited interpretability. Interpretability of machine

learning algorithms is a topic receiving significant focus [27]. This concept of interpretability attempts to balance predictability of models with the ability to provide insights to understand the basis of prediction. Interpretability allows experts in a given field to validate the methods and/or features applied within the algorithm.

A research gap exists for chatter detection based on easily understood machine learning techniques and audio data features (e.g., amplitude, frequency and power coefficients). Our study attempts to mitigate this gap by implementing a Random Forest (RF) algorithm with feature reduction to classify audio data features based on presence of chatter. RF is a "black box" method, but the approach is conducive to increased understanding. Experimental approaches using feature reduction increase this interpretability [28]. Our study applies a combination of RF and Recursive Feature Elimination (RFE) to address this research gap. Researchers have extensively utilized RF algorithms and RFE to increase the ease of understanding of results [29, 30]. However, to our knowledge, the research community has not yet applied this approach in the detection of chatter based on audio data features.

This is significant in the ability to leverage such an approach to provide feedback during the machining process. The data utilized in our research is approximately 2 s per machining operation. Rapid analysis through methods such as RF with RFE can provide real-time feedback on the stability of a machining operation. This can enable both a validation of expert machinist views as well as assist new machinists in the determination of cut stability.

2 Methodology

In this study we use a RF algorithm to classify machining audio data as stable or unstable (indicating moderate to severe chatter). RF is a "black box" approach that uses multiple decision trees to categorize based on features. This approach is capable of classification based on very large sets of features [31, 32]. In addition to RF, we use RFE to both reduce the number of features necessary for prediction and improve model performance. RFE is an iterative approach that removes the least important feature in classification and can improve predictor selection [29]. An advantage of using RF and RFE is the ease of understanding results. By leveraging this approach, we can convey both the predictive features and their relative importance.

Stability prediction within this research is solely based on the classification output of an RF classifier. Our approach extracts many features and then iteratively reduces the feature data set to both improve prediction accuracy and reduce processing requirements for feature extraction and prediction. This aligns with our vision of progression into

real-time classification using a portable device such as a laptop or smartphone.

2.1 Data description

The source data for this study is comprised of audio files collected by a single microphone during down-milling operations by a single machine at TU Wien, Vienna, Austria. We have categorized these audio files into three groups, as shown in Table 1. Specific details on the configuration of each down-milling operation, including spindle speed, axial depth, tooth passing frequency, and feed rate, are included in Table 2 (for set A only) and Table 8 in Appendix A (for all down-milling operations).

Sets A and A' have the same tool setup. Both set A and set A' tests are completed using a four-flute, 10-mm-diameter end mill that was clamped in a sensory tool holder. This tool holder included machined pockets that contained sensors and electronics. These pockets caused asymmetry in the dynamics between the two directions. Specifically, the dynamic response was stiffer in the direction without the pockets and more flexible in the direction aligned with the pockets. The feed per tooth is 0.1 mm/tooth, and the radial depth of cut is 2 mm. The cuts are down-milling. The machine tool is a DMG Mori DMU 75 monoBlock. The workpiece is Aluminum 6060-T66. All operations are carried out with flood coolant. The tool is re-assembled between set A and set A' using the same configuration. Set B utilizes a different tool setup. A stability lobe diagram for set A is presented in Fig. 2. This diagram was calculated using the stiff direction dynamics only. The disagreement between the stability lobe diagram and stable/unstable points occurs because the model does not incorporate the asymmetric rotating dynamics.

Set B tests are completed using a four-flute, 12-mm-diameter end mill. The feed per tooth is 0.1 mm/tooth, and the radial depth of cut is 3 mm. The cuts are down-milling. The machine tool is a DMG Mori DMU 75 monoBlock. The workpiece is Aluminum 6060-T66. All

Table 1 Audio file groups

% Group	Machine Similarity	TPF Classification	
		Stable	Unstable
Set A	Set A'	5	6
Set A'	Set A	5	12
Set B	N/A	4	3

Note: TPF classification represents physics-aware frequency analysis and classification. This classification uses a 50% threshold, meaning if the chatter frequency peak is greater than 50% of the peak of the TPF or its harmonics, the audio is classified as chatter

Table 2 Audio file machining parameters from set *A*

Audio Group	Spindle Speed	Axial Depth	Tooth Passing Frequency	Feed Rate	Classification
Set *A*	9000	4.0	600	3600	Stable
Set *A*	9000	5.2	600	3600	Stable
Set *A*	9000	6.1	600	3600	Unstable
Set *A*	9000	5.5	600	3600	Stable
Set *A*	9000	5.9	600	3600	Unstable
Set *A*	8510	6.9	567.3	3404	Unstable
Set *A*	8170	7.0	544.7	3268	Unstable
Set *A*	8700	6.3	580	3480	Stable
Set *A*	8700	6.7	580	3480	Stable
Set *A*	8700	7.2	580	3480	Unstable
Set *A*	8710	6.9	580.7	3484	Unstable

Note: TPF classification represents physics-aware frequency analysis and classification. This classification uses a 50% threshold, meaning if the chatter frequency peak is greater than 50% of the peak of the TPF or its harmonics, the audio is classified as chatter

operations are carried out with flood coolant. A physics-aware frequency analysis leveraging the machining spindle speed and the number of cutting teeth, resulting in the Tooth Passing Frequency (TPF), provides initial classification of audio files. Content at frequencies other than the TPF and its multiples indicates the potential for chatter. This information provides the classification of each audio file as either stable or unstable that we use as training data for the RF model.

Sets *A* and *A′* include all audio files collected during the machining execution in early 2021. Set *B* includes a subset of audio files collected during execution in late 2021. We selected specific audio files from the set *B* execution based on our ability to leverage common pre-processing techniques, aligned with those used with sets *A* and *A′*, to determine the time period between when the tool becomes fully engaged in the cut and starts to exit the cut.

2.2 Data pre-processing

To prepare the data for analysis, we pre-process each individual audio file using Python scripts [33] with the librosa [34, 35] and peakutils [36] packages. The first stage of pre-processing identifies the time period between the tool fully engaging in the cut and exiting the cut based on amplitude, percussion, and harmonic variations. Figure 3

Fig. 2 Set *A* stability lobe diagram based on the two stiff directions of the sensory tool holder. Stable points are represented by open circles. Unstable points are represented by x-marks. Specific data point parameters are detailed in Table 2. The presence of different stiffness directions causes inaccurate stability lobe diagrams
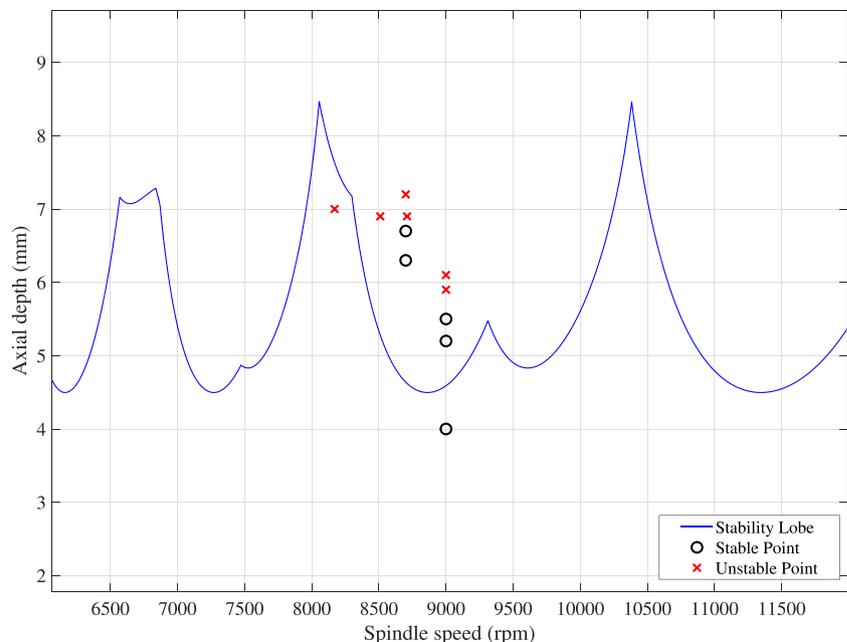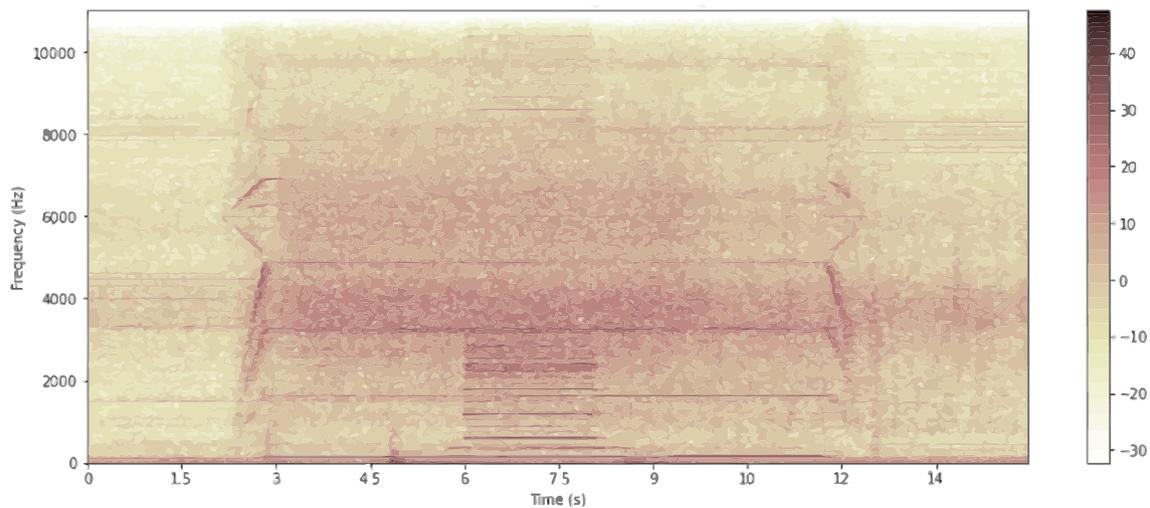
(a) Machining spectrogram.
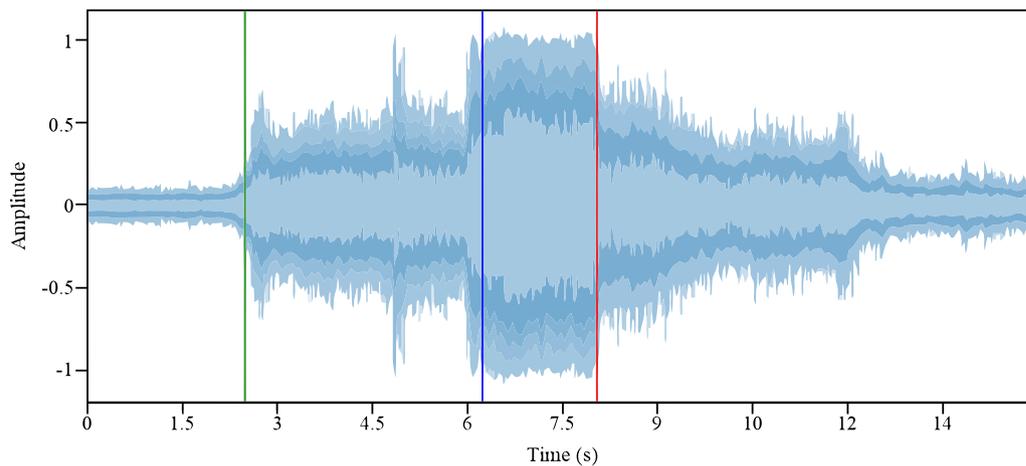
(b) Machining waveplot.

Fig. 3 Machining spectrogram and waveplot: (a) illustrates the frequency variation over the audio capture timeframe with the tool becoming fully engaged at approximately 6 s; (b) illustrates the amplitude over the audio capture timeframe with spindle startup noted by the vertical green line, the tool becoming fully engaged noted by the vertical blue line, and the tool starting to exit the cut noted by the vertical red line

provides spectrogram and waveplot images that illustrate this timeframe. The second stage of pre-processing extracts features for analysis. Figure 4 illustrates the extraction of peak magnitudes for the top 10 peak features.

The feature extraction results in a total of 152 features from each audio file. We group these features as spectral, harmonic, and peak features. Spectral features include 64 descriptive statistics within the categories of spectral centroids, spectral rolloff, and spectral bandwidth. Harmonic features include 24 descriptive statistics within the categories of harmonics, beat track, and perpetual shock. Peak features include 50 direct measurements from the ten peaks with greatest magnitude and 24 descriptive statistics derived from the direct measurements.

2.3 Model development

Our overall research approach includes feature pruning, feature elimination, and cross-group analysis. We present an overview of our approach in Fig. 5.

Feature pruning includes removal of non-predictive features and consolidation of features that were highly correlated. Some features have no variation across individual audio files within specific audio file sets, regardless of whether the audio feature represents a stable or unstable cut. These features are non-predictive. Other features are highly correlated and are redundant within individual audio file sets. We perform correlation analysis to identify non-predictive features and features correlated at or above 95%.
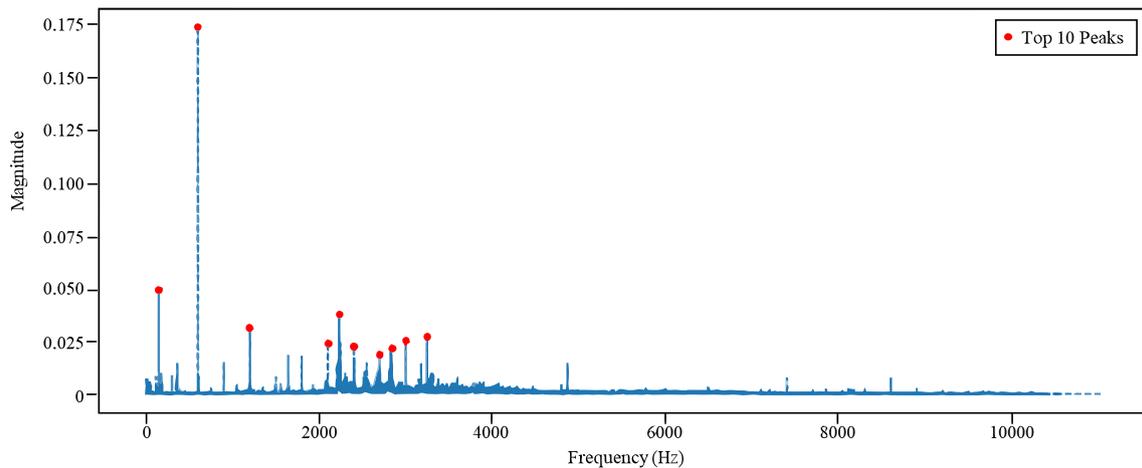
Fig. 4 Our approach further analyzes the machining timeframe to determine magnitude of the top 10 peak features

We then remove all non-predictive features and all but one feature from highly correlated groups. This results in a smaller set of features that have potential for prediction.

Feature elimination focuses on identifying the top features within categories of spectral, harmonic, and peak features. RFE identifies the least important feature within a given RF model and then removes this feature for subsequent iterations. We perform RFE through each feature category and note the top five features from each category.

Core to our approach is the utilization of Leave-One-Out Cross Validation (LOOCV). This method is effective for data sets with small numbers of samples. LOOCV utilizes one observation from a data set as a testing component and the remaining observations as training components. This generates a k-fold model with a fold size of one [37]. Figure 6 illustrates LOOCV data segmentation. LOOCV was utilized based upon the small number of samples.

2.4 Evaluation metrics

Our investigation includes multiple studies to determine the consistency of key features across data sets, predictability of RF models within data sets, and generalizability of RF models.

We will evaluate the consistency of key features across data sets through the cardinal number derived from the intersection of key features identified through RFE. The cardinal number represents the number of key features in common across sets. Additionally, we will explore the specific features included in that intersection.

We will evaluate RF models both within audio file groups (i.e., a single setup) and across audio file groups (i.e., multiple setups). The evaluation within audio file groups will provide insights into the ability to predict chatter within a single machining setup. The evaluation across audio groups

will provide insights into the generalizability of RF models across varying machining setups. Evaluation metrics for RF models will include Receiver Operating Characteristics (ROC) Area Under Curve (AUC), sensitivity, and specificity. The ROC curve is a plot with the true positive rate on the vertical axis and the false positive rate on the horizontal axis. Each axis has a range from 0 to 1. The AUC metric calculates the area under this curve. The maximum area possible is 1, and the minimum area possible is 0. Random guessing results in an AUC value of 0.5; values above 0.5 indicate greater predictability than random guessing. Equation (1) details the sensitivity calculation. Equation (2) details the specificity calculation.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (1)$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \quad (2)$$

AUC can be explained as the "probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example," providing both discrimination and consistency. This combination is advantageous over other metrics such as accuracy [38]. Therefore, the primary evaluation metric for our study is AUC.

3 Investigation

This paper includes four studies to determine the consistency of key features across data sets, predictability of RF models within data sets, and generalizability of RF models. These studies include use of the top five important features in each feature category. The quantity of top features was selected based on a review of overall feature importances
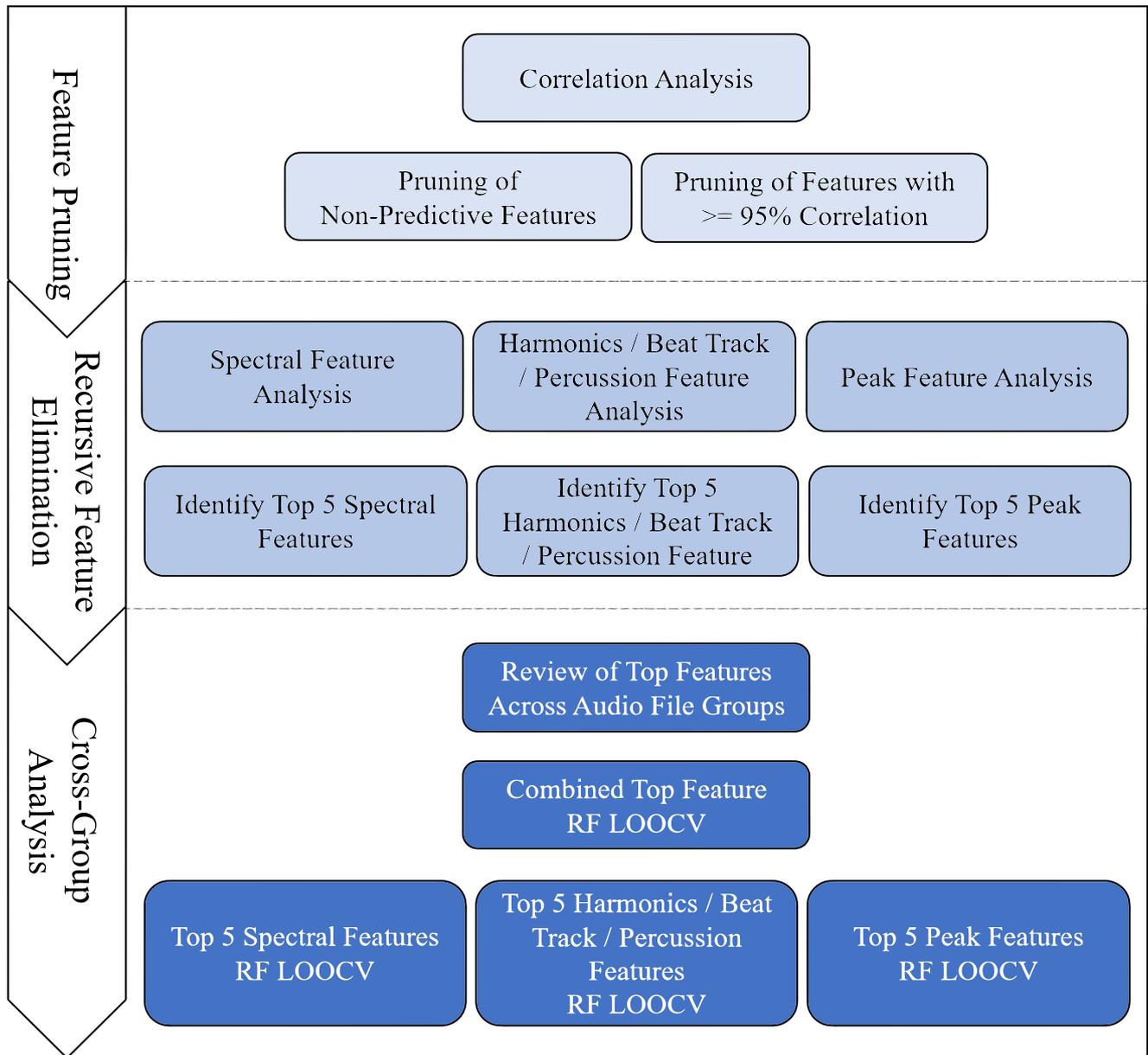
Fig. 5 The research approach includes feature pruning, recursive feature elimination, and cross-group analysis. Steps in this approach are provided with a generalized flow from top to bottom. The application of Random Forest Leave-One-Out Cross Validation is noted as RF LOOCV

and the AUC values across the three data sets. The peak AUC was typically achieved with between one and eight features included.

Study 1: Compare features across set $i \in \{A, A'\}$ to determine if there is consistency of important features in machining audio with the same tool setups.

Step 1: Evaluate top 5 important features, which we denote by set F_i^c, for set $i \in \{A, A'\}$ and feature category $c \in \{\text{spectral, harmonic, peak}\}$

Step 2: Determine intersection $F_A^c \cap F_{A'}^c$

Step 3: Determine cardinal number $n(F_A^c \cap F_{A'}^c)$

Study 2: Compare features across set $i \in \{A, A'\}$ and set $j \in \{B\}$ to determine if there is consistency of important features in machining audio with differing tool setups.

Step 1: Evaluate top 5 important features, which we denote by set F_i^c, for set $i \in \{A, A'\}$ and feature category $c \in \{\text{spectral, harmonic, peak}\}$

Step 2: Evaluate top 5 important features, which we denote by set F_j^c, for set $j \in \{B\}$ and feature category $c \in \{\text{spectral, harmonic, peak}\}$

Step 3: Determine intersection $F_A^c \cap F_B^c$

Step 4: Determine cardinal number $n(F_A^c \cap F_B^c)$

Step 5: Determine intersection $F_{A'}^c \cap F_B^c$

Step 6: Determine cardinal number $n(F_{A'}^c \cap F_B^c)$

Study 3: Evaluate RF model performance across set $i \in \{A, A'\}$ to determine if there is generalizability of models in machining audio with the same tool setups.

Step 1: Create RF models for set $i \in \{A, A'\}$ and feature category $c \in \{$spectral, harmonic, peak, top_15_combined$\}$

Step 2: Evaluate model performance within each set utilizing LOOCV, expressed as $AUC_{A,A}^c$ and $AUC_{A',A'}^c$, with the model generating set followed by evaluation set in the sub-script

Step 3: Evaluate model performance across sets $i \in \{A, A'\}$ utilizing LOOCV, expressed as $AUC_{A,A'}^c$ and $AUC_{A',A}^c$, with the model generating set followed by evaluation set in the sub-script

Study 4: Evaluate RF model performance across set $i \in \{A, A'\}$ and set $j \in \{B\}$ to determine if there is generalizability of models in machining audio with differing tool setups.

Step 1: Create RF models for set $i \in \{A, A'\}$ and feature category $c \in \{$spectral, harmonic, peak, top_15_combined$\}$

Step 2: Create RF models for set $j \in \{B\}$ and feature category $c \in \{$spectral, harmonic, peak, top_15_combined$\}$

Step 3: Evaluate model performance within each set utilizing LOOCV, expressed as $AUC_{B,B}^c$, with the model generating set followed by evaluation set in the sub-script

Step 4: Evaluate model performance across sets $i \in \{A, A'\}$ and set $j \in \{B\}$ utilizing LOOCV, expressed as $AUC_{A,B}^c$, $AUC_{A',B}^c$, $AUC_{B,A}^c$, and $AUC_{B,A'}^c$, with the model generating set followed by evaluation set in the sub-script

4 Results

4.1 Features and descriptive statistics

Feature extraction creates 152 features within the spectral, harmonic, and peak feature groups. Each feature group includes multiple sub-groups, and each sub-group includes multiple key features. Descriptive statistics further expand our understanding of these key features. Table 3 provides a summary of these feature groups, sub-groups, and key features. Tables and figures through the remainder of this document use abbreviations for individual descriptive statistics. For example, Bandwidth_4_Median represents the fourth order spectral bandwidth median and Centroids_Kurt

Fig. 6 The LOOCV method iterates through data selecting one sample as the test set while retaining other samples as the training set

represents spectral centroids kurtosis. Tables 9, 10, and 11 in Appendix B provide a listing of all key feature abbreviations noted in this paper.

4.2 Investigation results

Studies 1 and 2 provide insights into the consistency of key features across data sets. These studies utilize 10 iterations of feature ranking with randomly generated seeds. The results of these iterations are averaged for each step during RFE and the least important feature, based on mean importance, is removed. Our evaluation focuses on the 5 most important features in each feature category.

Studies 3 and 4 provide insights into the predictability of RF models within data sets and the generalizability of RF models across data sets. These studies utilize 10 iterations of RF models evaluation with randomly generated seeds. The results of these iterations produce mean and 95% confidence intervals for AUC, sensitivity and specificity.

4.2.1 Study 1 results

Study 1 results reveal a varying level of alignment of key features within sets A and A'. The cardinal number for spectral features is 0. The cardinal number for harmonic features is 1, with perpetual shock maximum present as a key feature in both sets. The cardinal number for peak features is 3, with peak magnitude skew, peak 2 normalized magnitude, and peak 3 normalized magnitude present as key features in both sets. Figure 7 details the top features for sets A and A' for each feature category.

4.2.2 Study 2 results

Study 2 results reveal a varying level of alignment of key features between sets A and A' and set B. The cardinal number for spectral features across sets A and B is 0. The cardinal number for spectral features across sets A' and B is 1, with spectral centroids bandwidth 2 minimum present as a key feature in both sets. The cardinal number for harmonic features across sets A and B is 0. The cardinal number for harmonics features across sets A' and B is 0. The cardinal

Table 3 Feature groups and key features

Feature Group	Feature Sub-group	Description
Spectral	Centroids	Centroids describe the center of mass for a sound. Key features related to spectral centroids include spectral centroids, spectral centroids delta, and spectral centroids accelerate.
	Rolloff	Rolloff is the frequency below which a set percentage of spectral energy exists.
	Bandwidth	Bandwidth provides the pth order spectral bandwidth. Key features include spectral bandwidth 2, spectral bandwidth 3, and spectral bandwidth 4.
Harmonic	Harmonics	Harmonics represent the harmonic class of sound extracted from an audio file through harmonic-percussive separation.
	Beat Track	Beat track describes the tempo of sound.
	Perpetual Shock	Perpetual shock represents the percussive class of sound extracted from an audio file through harmonic-percussive separation.
Peak	Peak Magnitude	Peak magnitude is the amplitude associated with specific audio peaks. Key features related to peak magnitude are associated with the top 10 frequency peaks in a sound file, such as peak 2 magnitude.
	Peak Frequency	Peak frequency is the frequency associated with specific audio peaks. Key features related to peak frequency are associated with the top 10 frequency peaks in a sound file, such as peak 2 frequency.
	Peak Normalized Magnitude	Peak normalized magnitude is a normalized amplitude associated with specific audio peaks. Key features related to peak normalized magnitude are associated with the top 10 frequency peaks in a sound file, such as peak 2 normalized magnitude.
	Peak Frequency Gap	Peak frequency gap is the frequency distance between a given peak and the highest amplitude peak. Key features related to peak frequency gaps are associated with the top 10 frequency peaks in a sound file, such as peak 2 frequency gap.

number for peak features across sets A and B is 1, with peak 3 normalized magnitude present as a key feature in both sets. The cardinal number for peak features across sets A' and B is 1, with peak 3 normalized magnitude present as a key feature in both sets. Figure 8 details the top features for set B in each feature category.

4.2.3 Study 3 results

Study 3 results reveal high predictability for RF models within data sets. Set A results are included in Table 4. For this set, all AUC values are above 0.95. Set A' results are included in Table 5. This set does not demonstrate as high of predictability across all feature groups, but many are highly predictive.

$AUC_{A,A}^{\text{peak}}$, which describes predictability of peak features within set A, is 1.000. $AUC_{A,A}^{\text{harmonics}}$, which describes predictability of harmonic features within set A, is 0.992.

$AUC_{A,A}^{\text{spectral}}$, which describes predictability of spectral features within set A, is 0.950. Set A's combined top 15 feature predictability, $AUC_{A,A}^{\text{top-15}}$, is 1.000.

$AUC_{A',A'}^{\text{spectral}}$, which describes predictability of spectral features within set A', is 0.997. $AUC_{A',A'}^{\text{peak}}$, which describes predictability of peak features within set A, is 0.935. $AUC_{A',A'}^{\text{harmonics}}$, which describes predictability of harmonic features within set A, trails at 0.669. Set A''s combined top 15 feature predictability, $AUC_{A',A'}^{\text{top-15}}$, is 0.983.

Study 3 results also reveal some generalizability for RF models across data sets with similar tool setups. Results for feature model generalizability are included in Table 7. Peak features provide the highest AUC values, with $AUC_{A,A'}^{\text{peak}}$ at 0.891 and $AUC_{A',A}^{\text{peak}}$ at 0.980. The combined top 15 features also provide AUC values that indicate generalizability, with $AUC_{A,A'}^{\text{top-15}}$ at 0.777 and $AUC_{A',A}^{\text{top-15}}$ at 0.777. Spectral and harmonics feature generalizability trails significantly.
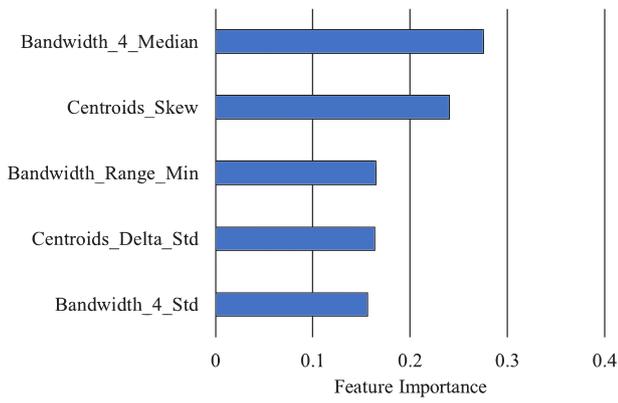
(a) Set A top five spectral features.
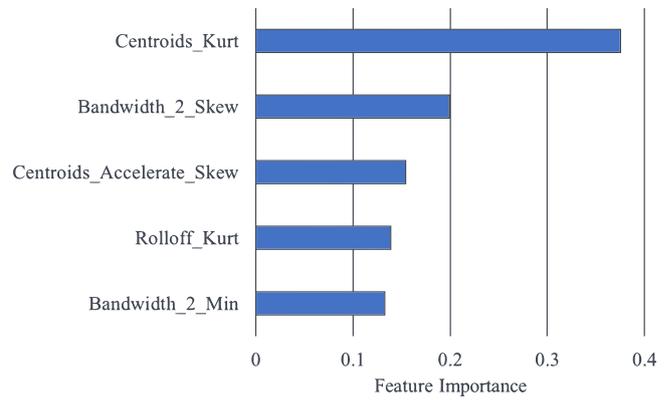
(b) Set A' top five spectral features.
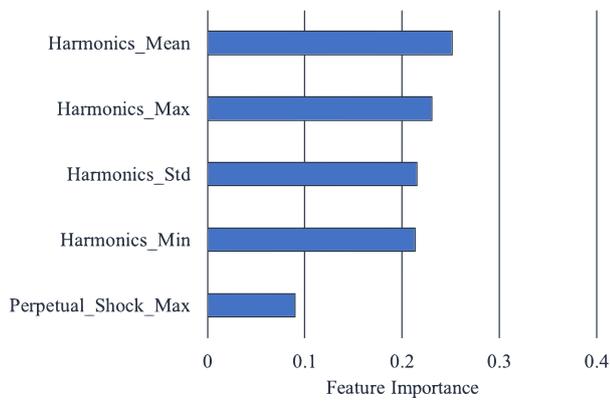
(c) Set A top five harmonic features.
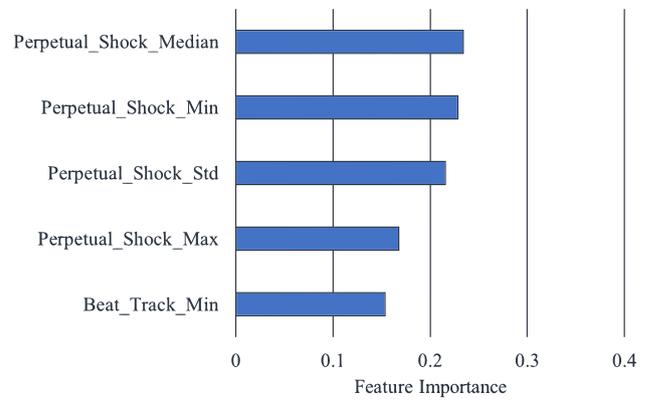
(d) Set A' top five harmonic features.
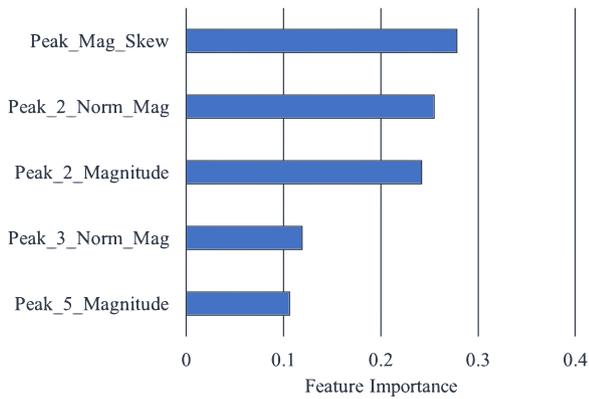
(e) Set A top five peak features.
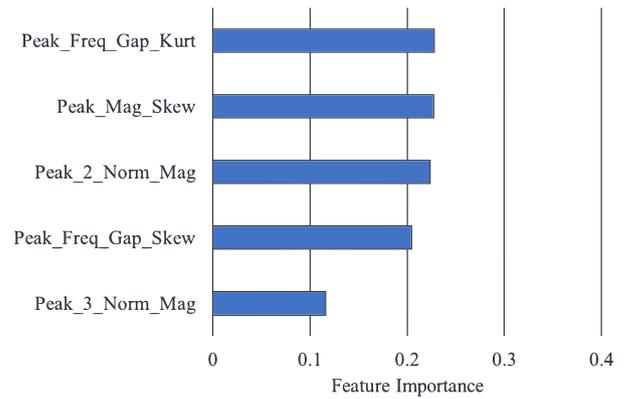
(f) Set A' top five peak features.

Fig. 7 Set A and A' feature importances derived from RFE include the following: (a) top five spectral features of set A; (b) top five spectral features of set A'; (c) top five harmonic features of set A; (d) top five harmonic features of set A'; (e) top five peak features of set A; (f) top five peak features of set A'. Each plot includes five key features with the relative importance indicated by the horizontal axis. The sum of all relative importances in each individual plot is equal to one

(a) Set B top five spectral features.

(b) Set B top five harmonic features.

(c) Set B top five peak features.
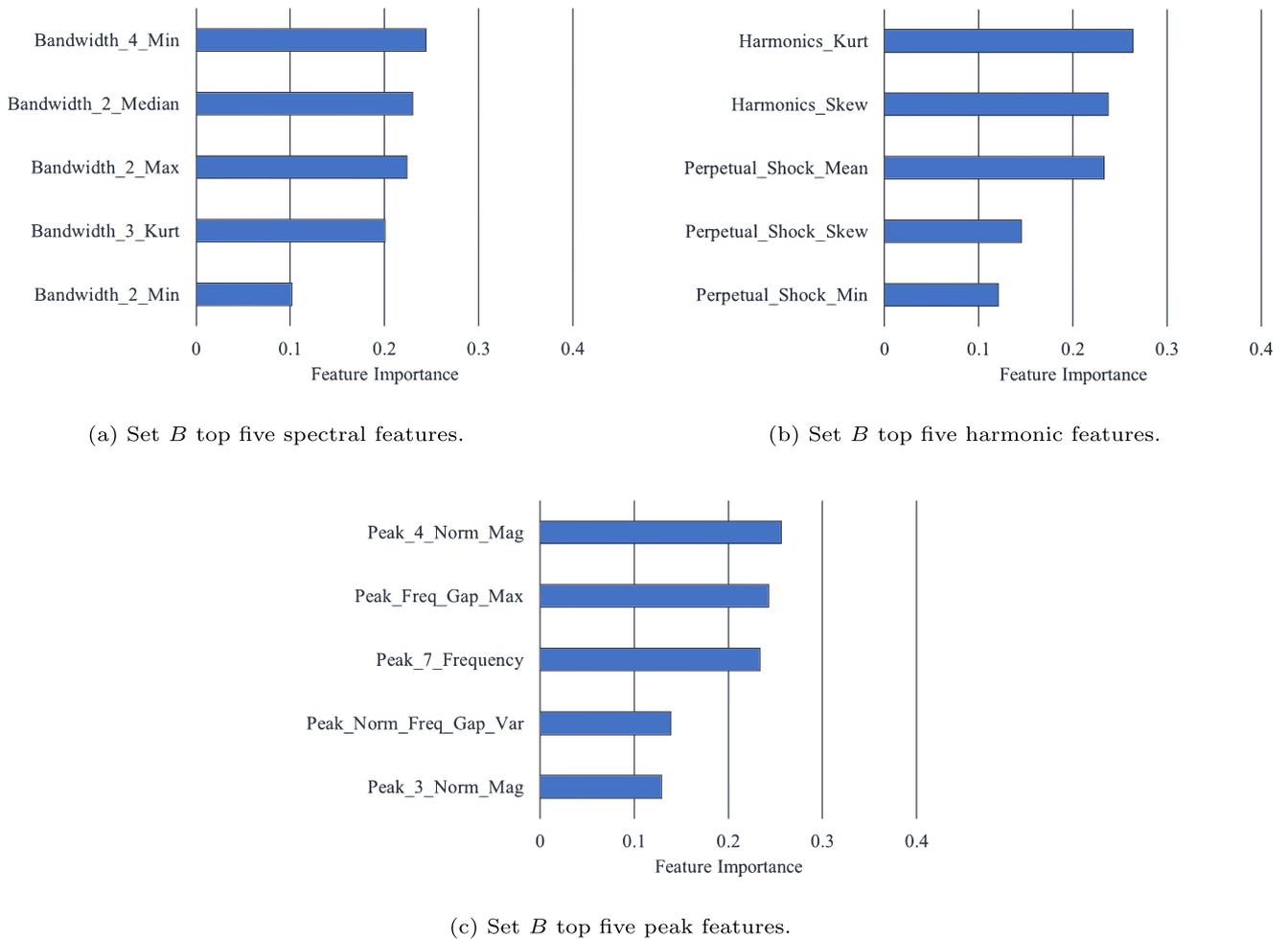
Fig. 8 Set B feature importances derived from RFE include the following: (a) top five spectral features of set B; (b) top five harmonic features of set B; (c) top five peak features of set B

4.2.4 Study 4 results

Study 4 results reveal modest predictability for RF models within data set B. These results are included in Table 6. $AUC_{B,B}^{\text{spectral}}$, which describes predictability of spectral features within set B, is 0.958. $AUC_{B,B}^{\text{peak}}$, which describes predictability of peak features within set B, is 0.717. $AUC_{B,B}^{\text{harmonics}}$, which describes predictability of harmonic

features within set B, is 0.667. Set B's combined top 15 feature predictability, $AUC_{B,B}^{\text{top\_15}}$, is 0.925.

Study 4 results also reveal modest generalizability for RF models with varying tool setups. These results are included in Table 7. Generalizability is noted within peak feature evaluations, where all AUC values are greater than 0.6. Spectral features provide inconsistent generalizability and harmonic features are not generalizable.

Table 4 Set A Random Forest model predictability

Model Features	AUC	Sensitivity	Specificity
All pruned features	0.982 ± 0.013	0.959 ± 0.047	0.950 ± 0.044
Top five spectral features	0.950 ± 0.024	0.963 ± 0.065	0.851 ± 0.067
Top five harmonic features	0.992 ± 0.008	0.978 ± 0.039	0.950 ± 0.044
Top five peak features	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Combined top 15 features	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Note: Evaluation metrics are based on a RF model using LOOCV with set A data. This evaluation includes 10 iterations using random seeds

Table 5 Set A' Random Forest model predictability

Model Features	AUC	Sensitivity	Specificity
All pruned features	0.888 ± 0.028	0.778 ± 0.049	0.847 ± 0.002
Top five spectral features	0.997 ± 0.004	1.000 ± 0.000	0.874 ± 0.030
Top five harmonic features	0.669 ± 0.009	0.611 ± 0.049	0.772 ± 0.013
Top five peak features	0.935 ± 0.003	0.972 ± 0.049	0.856 ± 0.002
Combined top 15 features	0.983 ± 0.011	1.000 ± 0.000	0.857 ± 0.000

Note: Evaluation metrics are based on a RF model using LOOCV with Set A' data. This evaluation includes 10 iterations using random seeds

Table 6 Set B Random Forest model predictability

Model Features	AUC	Sensitivity	Specificity
All pruned features	0.250 ± 0.106	0.489 ± 0.019	0.000 ± 0.000
Top five spectral features	0.958 ± 0.026	0.800 ± 0.000	1.000 ± 0.000
Top five harmonic features	0.667 ± 0.000	0.800 ± 0.000	1.000 ± 0.000
Top five peak features	0.717 ± 0.053	0.667 ± 0.000	1.000 ± 0.000
Combined top 15 features	0.925 ± 0.019	0.800 ± 0.000	1.000 ± 0.000

Note: Evaluation metrics are based on a RF model using LOOCV with Set B data. This evaluation includes 10 iterations using random seeds

Table 7 Random Forest model generalizability (AUC)

		Testing Set		
Features	Training Set	Set A	Set A'	Set B
All Pruned Features	Set A	–	0.805 ± 0.014	0.679 ± 0.023
	Set A'	0.813 ± 0.028	–	0.742 ± 0.055
	Set B	0.610 ± 0.047	0.798 ± 0.032	–
Top 5 Spectral Features	Set A	–	0.334 ± 0.014	0.771 ± 0.037
	Set A'	0.622 ± 0.025	–	0.300 ± 0.023
	Set B	0.460 ± 0.038	0.313 ± 0.046	–
Top 5 Harmonic Features	Set A	–	0.609 ± 0.012	0.467 ± 0.036
	Set A'	0.592 ± 0.025	–	0.283 ± 0.034
	Set B	0.468 ± 0.092	0.435 ± 0.026	–
Top 5 Peak Features	Set A	–	0.891 ± 0.011	0.800 ± 0.063
	Set A'	0.980 ± 0.019	–	0.613 ± 0.046
	Set B	0.652 ± 0.012	0.822 ± 0.009	–
Combined Top 15 Features	Set A	–	0.777 ± 0.011	0.800 ± 0.063
	Set A'	0.873 ± 0.022	–	0.475 ± 0.024
	Set B	0.610 ± 0.055	0.702 ± 0.058	–

Note: Evaluation metrics are based on a RF model trained on data from sets presented in the "Training Set" column and tested against data from sets in the "Testing Set" columns. These evaluations include ten iterations using random seeds

5 Discussion

As previously noted, RF models with RFE provide an approach that allows ease of understanding through clear indication of features utilized in classification. Our studies reveal significant variance between key features across data sets. This inconsistency indicates variability of key features in classification algorithms. Three of five key features align within the peak feature category across sets A and A'. This is the only instance where key feature alignment is above 40%.

Our studies show that RF models provide high predictability within distinct tool setups, with spectral and peak features providing the highest predictability within individual data sets. This predictability is evident from results within sets A and A', but somewhat less evident with set B. The discrepancy noted between sets A/A' and set B is not surprising based on the lower number of samples in set B. Set A has a total of eleven samples, while set B has seven. We can gather insights into the desired number of samples to achieve high predictability based on this difference. A minimum number of samples to drive predictability is likely between eight and eleven.

Our studies also demonstrate modest generalizability across similar tool setups. Peak features provide the highest generalizability across data sets. This generalizability is reduced when assessed across varying tool setups. Some of this loss of generalizability may be attributed to the low sample size within set B. However, it is unclear what impact the sample size has in comparison with tool setup.

6 Conclusions

In this study, we demonstrate that a RF classifier can effectively classify microphone-collected machining audio data as stable or unstable with relatively small sample sets. We believe a small initial sample of machinist classifications can result in high predictability for the duration of a tool-setup operation, resulting in automated chatter detection. This represents a low-cost and highly understandable approach with limited training data required. We also demonstrate that peak feature evaluation with RF algorithms has potential for generalizability across tool setups.

This work addresses easily understood classification of complete audio files, but does not address fully explainable methods or updated learning within an audio sample. Future work may expand upon our findings through the use of explainable artificial intelligence, such as SHapley Additive exPlanations (SHAP). This may provide insights into the variability of key features and will increase the ability of operators to comprehend the methods and basis of decision-making. Future work may also include methods that learn through the sequential processing of audio files, such as LSTM.

Appendix A. Audio file machining parameters

Table 8 Audio file machining parameters from all data sets

Audio Group	Spindle Speed	Axial Depth	Tooth Passing Frequency	Feed Rate	Classification
Set A	9000	4.0	600	3600	Stable
Set A	9000	5.2	600	3600	Stable
Set A	9000	6.1	600	3600	Unstable
Set A	9000	5.5	600	3600	Stable
Set A	9000	5.9	600	3600	Unstable
Set A	8510	6.9	567.3	3404	Unstable
Set A	8170	7.0	544.7	3268	Unstable
Set A	8700	6.3	580	3480	Stable
Set A	8700	6.7	580	3480	Stable
Set A	8700	7.2	580	3480	Unstable
Set A	8710	6.9	580.7	3484	Unstable
Set A'	8450	5.3	563.3	3380	Unstable
Set A'	8320	4.7	554.7	3328	Unstable
Set A'	8600	4.5	573.3	3440	Unstable
Set A'	8600	4.5	573.3	3440	Unstable
Set A'	8440	4.1	562.7	3376	Unstable
Set A'	8140	4.1	542.7	3256	Unstable
Set A'	8810	4.1	587.3	3524	Unstable
Set A'	7920	4.1	528	3168	Unstable
Set A'	8530	3.6	568.7	3412	Unstable
Set A'	8700	6.3	580	3480	Stable
Set A'	8240	3.7	549.3	3296	Unstable
Set A'	8680	3.3	578.7	3472	Stable
Set A'	8700	4.0	580	3480	Stable
Set A'	8710	4.4	580.7	3484	Unstable
Set A'	7570	5.5	504.7	3028	Unstable
Set A'	9000	4.0	600	3600	Stable
Set A'	9000	5.2	600	3600	Stable
Set B	18000	3.0	1200	7200	Stable
Set B	17000	3.0	1133.3	6800	Stable
Set B	16000	3.0	1066.7	6400	Stable
Set B	15000	3.0	1000	6000	Stable
Set B	16000	13.0	1066.7	6400	Unstable
Set B	17000	14.0	1133.3	6800	Unstable
Set B	17000	13.0	1133.3	6800	Unstable

Note: TPF classification represents physics-aware frequency analysis and classification. This classification uses a 50% threshold, meaning if the chatter frequency peak is greater than 50% of the peak of the TPF or its harmonics, the audio is classified as chatter

Appendix B. Key feature abbreviations

Table 9 Key spectral feature abbreviations

Key Feature Abbreviation	Key Feature
Centroids_Accelerate_Skew	Spectral centroids accelerate skew; describing the symmetry of the second derivative of the spectral centroids
Centroids_Delta_Std	Spectral centroids delta standard deviation; describing the distribution of the first derivative of the spectral centroids
Centroids_Kurt	Spectral centroids kurtosis; describing the tails of the distribution of spectral centroids
Centroids_Skew	Spectral centroids skew; describing the symmetry of the spectral centroids
Bandwidth_2_Median	Second order spectral bandwidth median
Bandwidth_2_Max	Second order spectral bandwidth maximum
Bandwidth_2_Min	Second order spectral bandwidth minimum
Bandwidth_2_Skew	Second order spectral bandwidth skew
Bandwidth_3_Kurt	Third order spectral bandwidth kurtosis
Bandwidth_4_Median	Fourth order spectral bandwidth median
Bandwidth_4_Min	Fourth order spectral bandwidth minimum
Bandwidth_4_Std	Fourth order spectral bandwidth standard deviation
Bandwidth_Range_Min	The difference between the fourth order spectral bandwidth and the second order spectral bandwidth minimum values
Rolloff_Kurt	Spectral rolloff kurtosis; describing the tails of the distribution of spectral rolloff, which is the frequency below which 85% of the total spectral energy exists

Table 10 Key harmonics feature abbreviations

Key Feature Abbreviation	Key Feature
Beat_Track_Min	Beat track minimum
Harmonics_Kurt	Harmonics kurtosis
Harmonics_Max	Harmonics maximum
Harmonics_Mean	Harmonics mean
Harmonics_Skew	Harmonics skew
Harmonics_Std	Harmonics standard deviation
Harmonics_Min	Harmonics minimum
Perpetual_Shock_Max	Perpetual shock maximum
Perpetual_Shock_Mean	Perpetual shock mean
Perpetual_Shock_Median	Perpetual shock median
Perpetual_Shock_Min	Perpetual shock minimum
Perpetual_Shock_Skew	Perpetual shock skew
Perpetual_Shock_Std	Perpetual shock standard deviation

Table 11 Key peak feature abbreviations

Key Feature Abbreviation	Key Feature
Peak_Freq_Gap_Kurt	Peak frequency gap kurtosis; peak frequency gap is the difference between the frequency peak with the highest magnitude and other top peaks
Peak_Freq_Gap_Max	Peak frequency gap max; peak frequency gap is the difference between the frequency peak with the highest magnitude and other top peaks
Peak_Freq_Gap_Skew	Peak frequency gap skew; peak frequency gap is the difference between the frequency peak with the highest magnitude and other top peaks
Peak_Mag_Skew	Peak magnitude skew
Peak_Norm_Freq_Gap_Var	Peak normalized frequency gap variance; peak normalized frequency gap is the difference between the normalized frequency peak with the highest magnitude and other normalized top peaks
Peak_2_Norm_Mag	Peak 2 normalized magnitude
Peak_2_Magnitude	Peak 2 magnitude
Peak_3_Norm_Mag	Peak 3 normalized magnitude
Peak_3_Magnitude	Peak 3 magnitude
Peak_4_Norm_Mag	Peak 4 normalized magnitude
Peak_7_Frequency	Peak 7 frequency

Data availability Audio files and derived data are not publicly available.

Code availability Code written in this study is not publicly available.

Declarations

Ethics approval The authors claim that there are no ethical issues involved in this research.

References

1. Taylor F (1906) On the art of cutting metals. Am Soc Mech Eng Proc Suppl 28(2):3–29
2. Tobias S, Fishwick W (1958) Theory of regenerative machine tool chatter. The Engineer 205(7):199–204
3. Schmitz T, Smith K (2019) Machining dynamics: frequency response to improved productivity, 2nd edn. Springer, New York. https://doi.org/10.1007/978-3-319-93707-6
4. Altintaş Y, Budak E (1995) Analytical prediction of stability lobes in milling. CIRP Ann Manuf Technol 44(1):357–362. https://doi.org/10.1016/S0007-8506(07)62342-7
5. Szalai R, Stepan G (2006) Lobes and lenses in the stability chart of interrupted turning. J Comput Nonlinear Dyn 1(3):205–211. https://doi.org/10.1115/1.2198216
6. Mann B, Edes B, Easley S, Young K, Ma K (2008) Chatter vibration and surface location error prediction for helical end mills. Int J Mach Tools Manuf 48(3-4):350–361. https://doi.org/10.1016/j.ijmachtools.2007.10.003
7. Duncan G, Kurdi M, Schmitz T, Snyder J (2006) Uncertainty propagation for selected analytical milling stability limit analyses. 34: 17–24
8. Liu Y, Li T, Liu K, Zhang Y (2016) Chatter reliability prediction of turning process system with uncertainties. Mech Syst Sig Process 66-67:232–247. https://doi.org/10.1016/j.ymssp.2015.06.030
9. Wang W, Wan M, Zhang W, Yang Y (2022) Chatter detection methods in the machining processes: a review. J Manuf Process 77:240–259. https://doi.org/10.1016/j.jmapro.2022.03.018
10. Yao Z, Mei D, Chen Z (2010) On-line chatter detection and identification based on wavelet and support vector machine. J Mater Process Technol 210(5):713–719. https://doi.org/10.1016/j.jmatprotec.2009.11.007
11. Yang K, Wang G, Dong Y, Zhang Q, Sang L (2019) Early chatter identification based on an optimized variational mode decomposition. Mech Syst Sig Process 115:238–254. https://doi.org/10.1016/j.ymssp.2018.05.052
12. Dun Y, Zhu L, Yan B, Wang S (2021) A chatter detection method in milling of thin-walled TC4 alloy workpiece based on auto-encoding and hybrid clustering. Mech Syst Sig Process 158:107,755. https://doi.org/10.1016/j.ymssp.2021.107755
13. Diniz A, Liu J, Dornfeld D (1992) Correlating tool life, tool wear and surface-roughness by monitoring acoustic-emission in finish turning. Wear 152(2):395–407. https://doi.org/10.1016/0043-1648(92)90135-U
14. Beggan C, Woulfe M, Young P, Byrne G (1999) Using acoustic emission to predict surface quality. Int J Adv Manuf Technol 15(10):737–742. https://doi.org/10.1007/s001700050126
15. Kishawy H, Hegab H, Umer U, Mohany A (2018) Application of acoustic emissions in machining processes: analysis and critical review. Int J Adv Manuf Technol 98(5-8):1391–1407. https://doi.org/10.1007/s00170-018-2341-y
16. Yang RY, Rai R (2019) Machine auscultation: enabling machine diagnostics using convolutional neural networks and large-scale machine audio data. Adv Manuf 7(2):174–187. https://doi.org/10.1007/s40436-019-00254-5
17. Aslan D, Altintas Y (2018) On-line chatter detection in milling using drive motor current commands extracted from CNC. Int J Mach Tools Manuf 132:64–80. https://doi.org/10.1016/j.ijmachtools.2018.04.007
18. Szydlowski M, Powalka B (2012) Chatter detection algorithm based on machine vision. Int J Adv Manuf Technol 62(5-8):517–528. https://doi.org/10.1007/s00170-011-3816-2
19. Schmitz T, Medicus K, Dutterer B (2002) Exploring once-per-revolution audio signal variance as a chatter indicator. Mach Sci Technol 6(2):215–233. https://doi.org/10.1081/MST-120005957
20. Bleicher F, Ramsauer C, Oswald R, Leder N, Schoerghofer P (2020) Method for determining edge chipping in milling based on tool holder vibration measurements. CIRP Ann Manuf Technol 69(1):101–104. https://doi.org/10.1016/j.cirp.2020.04.100
21. Sener B, Serin G, Gudelek MU, Ozbayoglu A, Unver H (2020) Intelligent chatter detection in milling using vibration data features and deep multi-layer perceptron. In: 2020 IEEE international conference on big data, IEEE international conference on big data, pp 4759–4768. https://doi.org/10.1109/BigData50022.2020.9378223
22. Sestito G, Venter G, Barros R, Kandice S, Rodrigues A, da Silva M (2022) In-process chatter detection in micro-milling using acoustic emission via machine learning classifiers. International Journal of Advanced Manufacturing Technology. https://doi.org/10.1007/s00170-022-09209-w
23. Jiang Y, Zhang C (2006) Hybrid HMM/SVM method for predicting cutting chatter. In: 3rd international symposium on precision mechanical measurements, vol 6280. International Society for Optics and Photonics, pp 396–403. https://doi.org/10.1117/12.716150
24. Wang R, Song Q, Liu Z, Ma H, Gupta M, Liu Z (2021) A novel unsupervised machine learning-based method for chatter detection in the milling of thin-walled parts. Sensors 21(17). https://doi.org/10.3390/s21175779
25. Vashisht R, Peng Q (2021) Online chatter detection for milling operations using LSTM neural networks assisted by motor current signals of ball screw drives. J Manuf Sci Eng Trans ASME 143(1). https://doi.org/10.1115/1.4048001
26. Kvinevskiy I, Bedi S, Mann S (2020) Detecting machine chatter using audio data and machine learning. Int J Adv Manuf Technol 108(11-12):3707–3716. https://doi.org/10.1007/s00170-020-05571-9
27. Carvalho D, Pereira E, Cardoso J (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8). https://doi.org/10.3390/electronics8080832
28. Wang X, Guo B, Shen Y, Zhou C, Duan X (2019) Input feature selection method based on feature set equivalence and mutual information gain maximization. IEEE Access 7:151,525–151,538. https://doi.org/10.1109/ACCESS.2019.2948095
29. Gregorutti B, Michel B, Saint-Pierre P (2017) Correlation and variable importance in random forests. Stat Comput 27(3):659–678. https://doi.org/10.1007/s11222-016-9646-1
30. Sun D, Wen H, Wang D, Xu J (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology 362. https://doi.org/10.1016/j.geomorph.2020.107201
31. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
32. Speiser J, Miller M, Tooze J, Ip E (2019) A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 134:93–101. https://doi.org/10.1016/j.eswa.2019.05.028

33. Van Rossum G, Drake F (2009) Python 3 Reference Manual (CreateSpace, Scotts Valley CA)
34. McFee B, Raffel C, Liang D, Ellis D, McVicar M, Battenberg E, Nieto O (2015) librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in science conference, vol 8. https://doi.org/10.25080/Majora-7b98e3ed-003
35. Librosa. https://doi.org/10.5281/zenodo.6097378. https://librosa.org/
36. Peakutils. https://doi.org/10.5281/zenodo.887917. https://pypi.org/
37. Wong T (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn 48(9):2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009
38. Huang J, Ling C (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17(3):299–310. https://doi.org/10.1109/TKDE.2005.50

Affiliations

Sam St. John[1] · Matthew Alberts[1] · Jaydeep Karandikar[2] · Jamie Coble[3] · Bradley Jared[4] · Tony Schmitz[2,4] · Christoph Ramsauer[5] · David Leitner[5] · Anahita Khojandi[1] (ID)

Sam St. John
sstjohn3@vols.utk.edu

Matthew Alberts
malberts@vols.utk.edu

Jaydeep Karandikar
karandikarjm@ornl.gov

Jamie Coble
jcoble1@utk.edu

Bradley Jared
bhjared@utk.edu

Tony Schmitz
tony.schmitz@utk.edu

Christoph Ramsauer
ramsauer@ift.at

David Leitner
leitner@ift.at

[1] Department of Industrial and Systems Engineering, University of Tennessee, 851 Neyland Drive, Knoxville, 37996, TN, USA

[2] Manufacturing Science Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, 37830, TN, USA

[3] Department of Nuclear Engineering, University of Tennessee, 863 Neyland Drive, Knoxville, 37996, TN, USA

[4] Department of Mechanical, Aerospace, and Biomedical Engineering, University of Tennessee, 1512 Middle Drive, Knoxville, 37996, TN, USA

[5] Institute of Production Engineering and Photonic Technologies, TU Wien, Vienna, Austria