Bearing Fault Type and Severity Prediction

-Bhargavi Jadhav

Problem Statement and Objective:

In safety-critical industrial machinery, bearing faults are a leading cause of mechanical failures, resulting in downtime, reduced productivity, and potential hazards. Early detection of bearing faults, along with severity assessment, is essential for implementing predictive maintenance strategies.

Traditional methods are often inadequate in identifying early-stage faults or quantifying severity. Hence, a data-driven approach using machine learning is required to enhance and automate the diagnosis.

In Scope:

1. Extraction of time-domain and frequency-domain features from vibration signals.
2. Develop a robust classification system to identify the type of fault (Inner Race, Outer Race, Ball Defect).
3. Predicting the severity of the fault (Small, Medium, Large).

Approach:

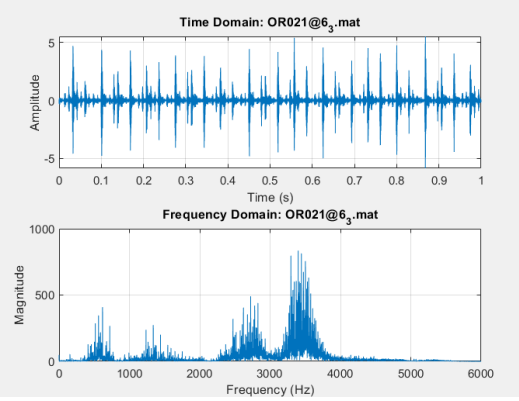
1. Data Preparation:
   i) The information about all the .mat files from the CWUR_data folder was stored in fileList. These files contain vibration data from the bearing at a sampling frequency of 12000.
   ii) The variables inside the .mat files were listed, and the Drive End (DE) vibration signal was selected (sensor placed near the motor side of the bearing).
   iii) The signal was then trimmed to 1 second to ensure uniform length.
   iv) The following **Time-Domain features** were extracted from the signal:
      (1) Root Mean Square (RMS): Indicates power in the signal. Higher RMS usually means higher vibration energy, often due to faults.
      (2) Skewness: Indicates asymmetry of the signal. This can suggest bearing damage.
      (3) Kurtosis: Tailedness or Peakiness. This feature is sensitive to impacts or shocks.
      (4) Mean: Average value of the signal. Not very sensitive to faults.
      (5) Standard Deviation: Indicates signal variability. Higher in faulty signals.
   v) **Fast Fourier Transform (FFT)** was used to convert the time-domain signal into the frequency domain. **Y** is the magnitude spectrum; **f** is the corresponding frequency array.
      (1) Peak Frequency: Frequency with highest amplitude. May correspond to defect frequencies.
      (2) Spectral Centroid: Center of mass of spectrum.
      (3) Spectral Spread: Variance around the centroid. Higher for complex, faulty signals.
      (4) Spectral Flatness: Flatness of the spectrum.
   vi) Fault type and fault severity were labelled based on the filenames
   vii) All the extracted features and labels were combined into one table.

viii) The table was saved to a CSV file and read from there:

A	B	C	D	E	F	G	H	I	J	K	
RMS	Skewness	Kurtosis	Mean	StdDev	PeakFreq	SpectralCentroid	SpectralSpread	SpectralFlatness	FaultType	Severity	FileName
0.138266	-0.02481	2.964506	0.015263	0.137426	3365	3000	2.45E-12	0.29906512	Ball	Small	B007_0.mat
0.138471	0.017413	2.983638	0.004209	0.138413	3478	3000	1.82E-12	0.299095748	Ball	Small	B007_1.mat
0.144331	-0.00617	2.724532	0.004849	0.144255	3432	3000	1.04E-12	0.323563911	Ball	Small	B007_2.mat
0.154805	0.035955	2.832997	0.003428	0.154773	3380	3000	8.60E-13	0.288017764	Ball	Small	B007_3.mat
0.176974	0.083109	10.43992	0.005013	0.176911	1437	3000	9.22E-13	0.367925013	Ball	Medium	B014_0.mat
0.147177	0.064147	7.840759	0.004555	0.147113	360	3000	1.19E-12	0.325250557	Ball	Medium	B014_1.mat
0.136895	0.013883	6.283448	0.00483	0.136816	360	3000	1.13E-12	0.357715696	Ball	Medium	B014_2.mat
0.125277	-0.06566	8.821625	0.004555	0.125199	360	3000	9.23E-13	0.358152824	Ball	Medium	B014_3.mat
0.115357	0.052484	3.591419	0.014136	0.114492	3208	3000	8.25E-13	0.41183301	Ball	Large	B021_0.mat
0.151414	-0.10676	9.965915	0.0043	0.15136	3385	3000	2.62E-12	0.374278343	Ball	Large	B021_1.mat
0.105022	-0.01419	3.345105	0.004539	0.104928	3345	3000	8.93E-13	0.436782557	Ball	Large	B021_2.mat
0.116355	0.042638	3.051508	0.004572	0.11627	3300	3000	9.26E-13	0.403001485	Ball	Large	B021_3.mat
0.289302	0.126775	5.632106	0.015136	0.288918	3587	3000	3.53E-12	0.363103422	InnerRace	Small	IR007_0.mat
0.293137	0.159998	5.385724	0.005782	0.293092	3539	3000	8.42E-13	0.370265382	InnerRace	Small	IR007_1.mat
0.29781	0.083936	5.536653	0.004347	0.297791	2862	3000	2.67E-12	0.346437029	InnerRace	Small	IR007_2.mat
0.312346	-0.01399	5.219968	0.004556	0.312326	2507	3000	1.78E-12	0.295643901	InnerRace	Small	IR007_3.mat

2. EDA:

i) A plot for a signal is drawn in order to understand the nature of the time-domain waveform. Similarly, the other graph shows how energy is distributed across the frequencies. This also helps to visualise the peak frequency, an important feature in the analysis.



ii) Feature Correlation Heatmap is plotted in order to calculate the correlation coefficients between all pairs of features.

Insights:

- SpectralCentroid was dropped in this because it remains constant across the samples.
- StdDev is perfectly correlated with RMS; similarly, SpectralFlatness is also highly correlated with RMS and StdDev. Hence, StdDev and SpectralFlatness are dropped in further analysis to avoid overfitting and redundancy.

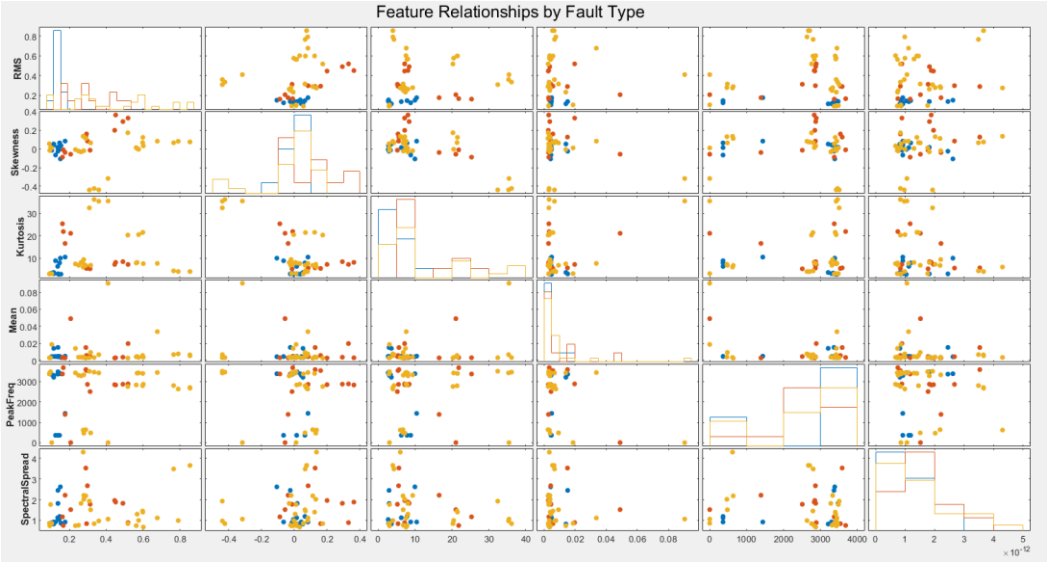
iii) Pairwise Scatter Plot and Feature distributions based on boxplots:

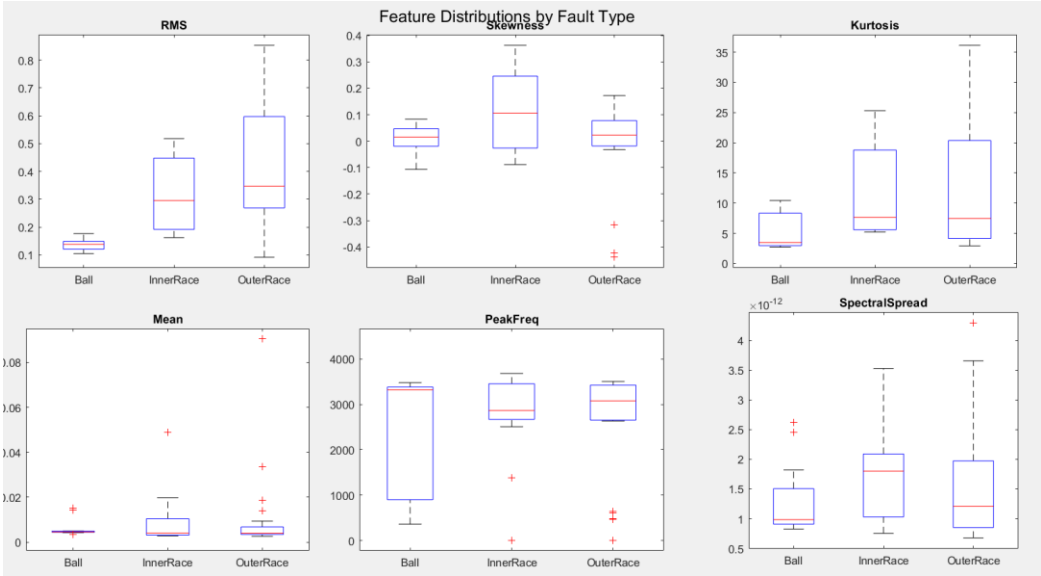
(1) Based on Fault Type (yellow=outer race, red= inner race, blue= ball) :

(a) Ball Fault has tightly clustered, lowered RMS values. Inner and outer race faults induce more vibration power.

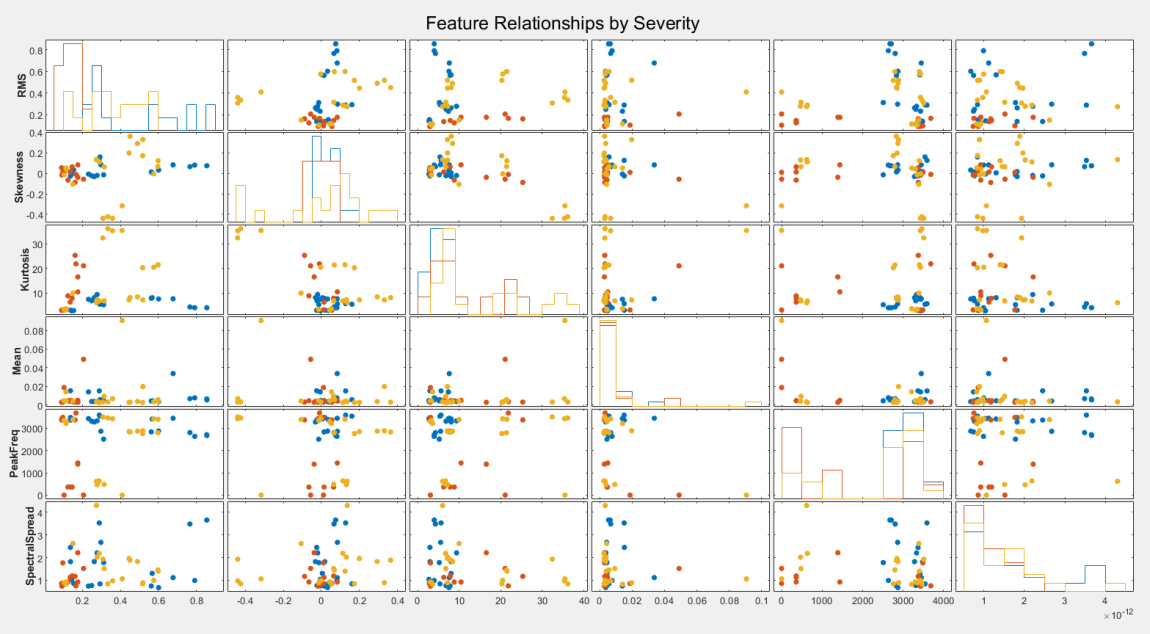
(b) Skewness is mostly clustered around 0, though it may be used to distinguish Inner Race (positive skew) faults from Ball and Outer Race faults that have skewness near-zero.

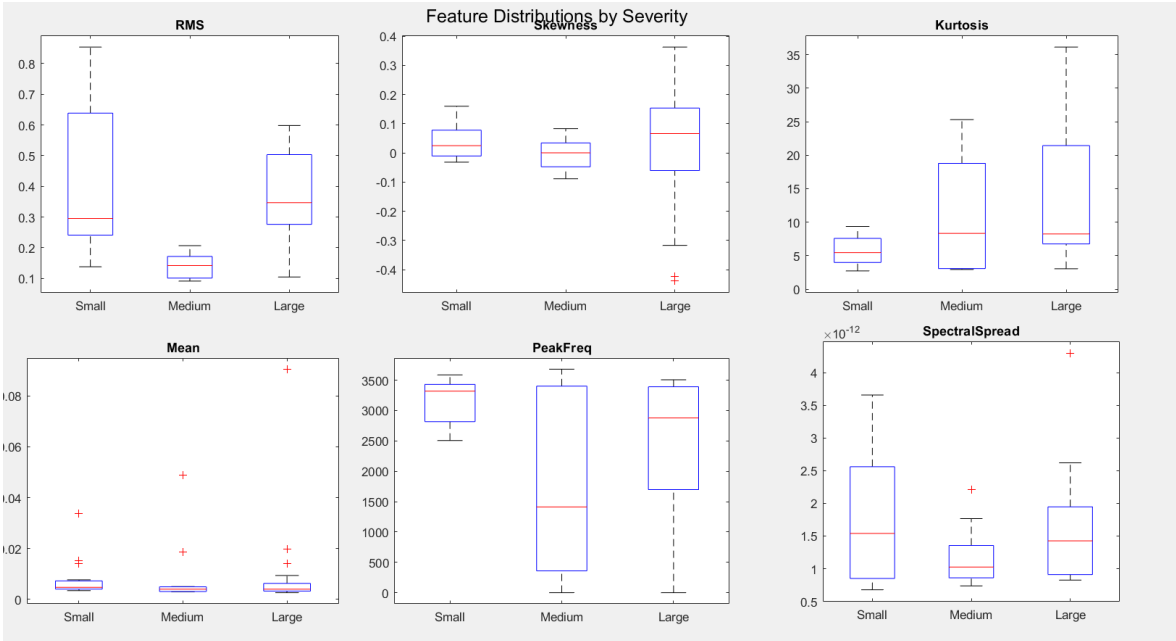
(c) Ball Faults can be associated with lower kurtosis values (~3-10), while much higher kurtosis values (up to 25-35) can be associated with Race Faults.
(d) The mean does not appear to be a strong differentiator between the faults.
(e) Most values of Peak Frequencies are clustered around 3000-3500 Hz for all types of faults. Ball Faults also show a wide spread, including lower frequencies as well.
(f) Race Faults show a wide spread of Spectral Spread as compared to Ball Faults.





(2) Based on Severity (yellow=large, red= medium, blue= small):
(a) Lower severity points (blue) are clustered more at lower values of RMS and dispersed towards higher values.
(b) All severities have Skewness centred around 0 with slight variation. No clear trend is observed.
(c) Higher values of Kurtosis are associated with higher severity, and low severity points are clustered towards lower values of Kurtosis.
(d) All the values are concentrated around the 0-mean value. This implies that the mean might have little contribution to the classifier.
(e) Points are mostly clustered towards higher values around 3000-4000 Hz. This can be the range where the natural frequency of the bearing lies. Some of the large and medium severity points can also be seen at lower frequencies. A possible reason could be that these faults started off with prominent peaks at higher frequencies, but as they evolved, peaks emerged at lower frequencies due to mechanical degradation spreading around.
(f) Distribution of Spectral Spread is more spread out in higher severity classes, while it is more clustered in lower ones.





3. Test-train split:
   For both the classifiers test data was split as 20% of the total data.

4. Model Training:
   (1) Fault Type prediction model:
   Multiple Classification models (Decision tree, k-NN, SVM, Bagged Tree) were evaluated using 5-fold cross-validation and checked for their accuracy. Based on these, the Bagged Tree model was chosen, and later, hyperparameter tuning was done to find a set of parameters that minimized the CV loss. The Bagged Tree model generally performs best in tabular data with relatively small datasets, as in this case.
   The Bagged Tree model uses ensemble learning to combine weak learners (trees) into a strong learner.

```
Model to predict severityDecision Tree CV Accuracy:66.67%
k-NN CV Accuracy:47.62%
SVM CV Accuracy:59.52%
```

```
Best estimated feasible point (according to models):
    NumLearningCycles    MinLeafSize    MaxNumSplits
    _________________    ___________    ____________

           486                1             11

Estimated objective function value = 0.2938
Estimated function evaluation time = 8.6681

Bagged Trees CV Accuracy:76.19%
```

(2) Severity prediction model:

A similar approach was used to train the model, except for hyperparameter tuning of the bagged tree model (may not be required).

```
Decision Tree CV Accuracy:66.67%
k-NN CV Accuracy:42.86%
SVM CV Accuracy:59.52%
Bagged Trees CV Accuracy:80.95%
```

5. Testing and understanding the performance:

Both models were tested on the test dataset and later evaluated on the following matrices:

(1) Fault Type prediction:

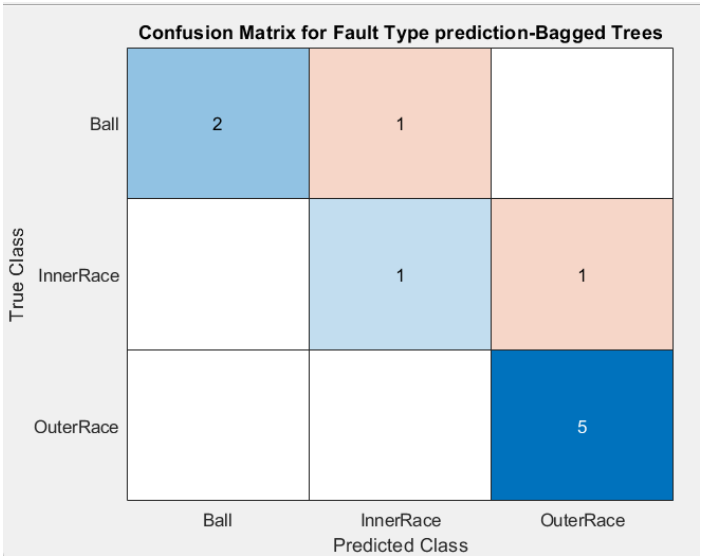
a) Accuracy:

It gives the percentage of correct predictions made by the model out of the total predictions.

```
Test Accuracy (Bagged Trees for fault type): 80.00%
```

b) Confusion matrix:

It was used to get a more detailed view of how predictions fell into each category, comparing the predicted and true classes.



c) Classification report of Precision, Recall, and F1 values

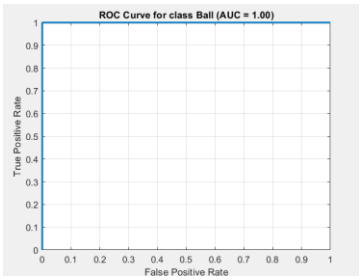
Precision	Recall	F1
0.66667	1	0.8
0.5	0.5	0.5
1	0.83333	0.90909

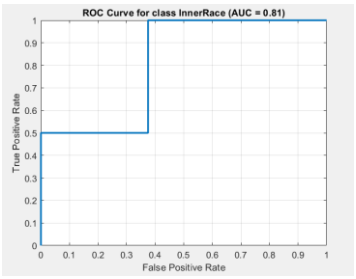
Precision gives the percentage of samples out of the predicted samples that were actually of that class. E.g., out of 3 Ball Faults predicted, 2 were true positives while 1 was a false positive. Hence, the precision for Ball Fault comes out to be 0.6667.

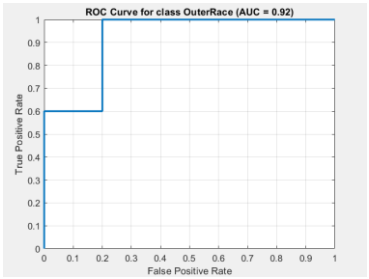
Recall gives the percentage of classes that were correctly predicted out of the actual samples. E.g., out of 6 Outer Race faults predicted, 5 were True positives. Hence recall=5/6=0.8333

F1 score gives the Harmonic mean of precision and recall values for a better understanding of how efficient the model is.

d) ROC-AUC curve:

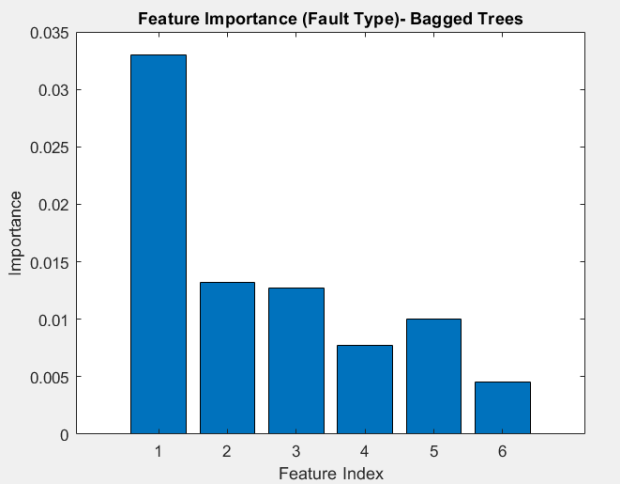






From the above values, it can be said that the model performed well overall for the Ball and Outer Race faults, but weakly for the Inner Race Faults. This could possibly be because of the small dataset that was available for training and testing.

e) Feature importance:



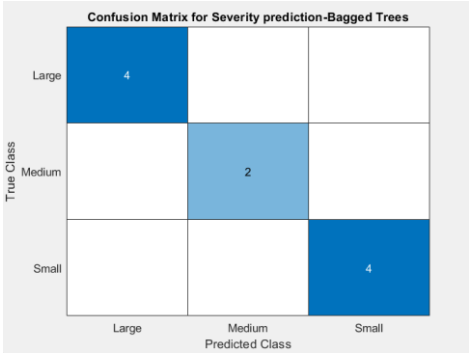
(2) Severity Prediction:

The same matrices were used to evaluate the severity prediction model:

a) Accuracy:

```
Test Accuracy (Bagged Trees for severity): 100.00%
```

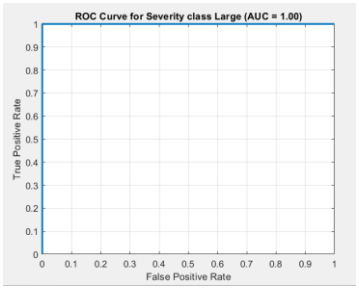
b) Confusion matrix:

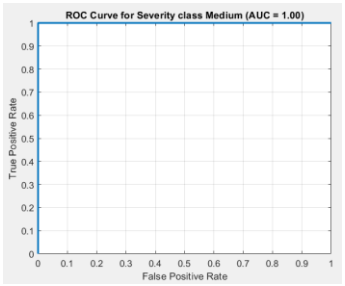


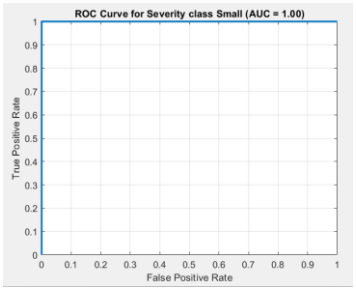
c) Classification report:

Precision	Recall	F1
1	1	1
1	1	1
1	1	1

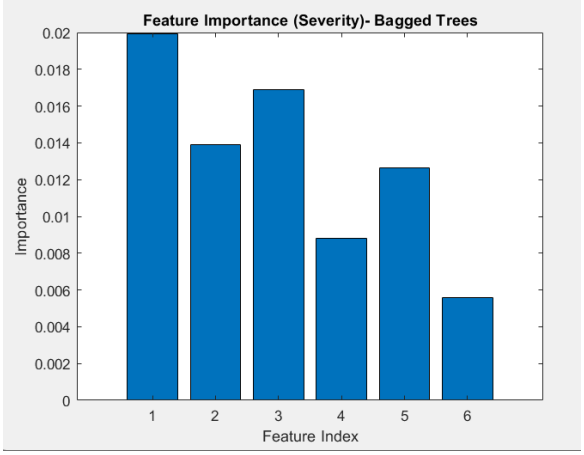
d) ROC-AUC curve:







e) Feature importance:



The severity classification model had around 80% cross-validation accuracy and achieved 100% accuracy on the test set. This could be attributed to the dataset being small or the test set being more representative of the training set.

6. Saving the models for future use:
   The models were saved in .mat format files and later used in functions for predictions.
   e.g.:

```
% Example feature row from your dataset:
sample_features = [0.14, 0.017, 2.6836, 0.0062, 3578, 1.82e-12];
```

   Prediction:

```
Predicted Fault Type: Ball
Predicted Fault Severity: Small
```

7. Conclusion and Future Scope:
   The bearing fault classification project involved building two machine learning models- one for fault type detection and the other for fault severity prediction- using vibration signal features. Overall, both models performed effectively; however, in order to deploy them in the real world, it is important to expand the dataset to reduce the risk of overfitting and make the models more reliable.