Improving Trust via XAI and Pre-Processing for Machine Learning of Complex Biomedical Datasets

Brandon D. Hines, Douglas A. Talbert, and Steven R. Anton
Dynamic and Smart Systems Laboratory (DSSL)
Tennessee Tech University
Cookeville, TN, USA
{bdhines42, dtalbert, santon}@tntech.edu

Abstract
Complex datasets hold a special place among engineers as the engineering community seeks to solve some of the world’s most difficult problems, but with complexity, comes difficulty in analysis and interpretation. Machine learning seeks to solve this problem by aiding in the analysis of these complex datasets, but the implementation of machine learning introduces an entirely new problem, that of transparency. Machine learning often results in models that perform the assigned task quite well, but the issue lies in the often black-box nature of machine learning models. Often in the engineering community, it can be difficult to trust black-box machine learning models to make crucial determinations within a complex dataset. This study explores this issue by expanding on previously conducted work in which a complex biomedical dataset was generated to explore the problem of persistent pain experienced after receiving a total knee replacement (TKR). Previous work saw the generation of simulated TKRs with varying levels of damage, and structural health monitoring (SHM) techniques were used to collect measurements on the samples with the hope of identifying a method to detect the presence of damage. Machine learning was then introduced as a technique to classify the varying levels of damage present in the dataset. This study explores the implementation of explainable artificial intelligence (XAI) as well as data processing techniques as a method to combat the black-box nature of machine learning models and generate more trustworthy models with respect to domain knowledge.

Introduction
Since its creation in 1890, Total Knee Arthroplasty (TKA) is a medical procedure that has been greatly optimized and widely adopted by the medical community (Eynon-Lewis, et al, 1992). The practice has grown into one of the most common medical procedures in the country. In the year 2010, an estimated 4.7 million people in the United States were living with a total knee replacement (TKR) (Kremers, et al, 2015). This number is likely higher today due to continuous population increase.

With the optimization of the procedure, it is shown that the procedure sees a high rate of success with 82% of TKRs still functioning properly after 25 years (Evans, et al, 2019). Though this is a high rate of success, approximately 20% of patients report experiencing persistent pain after the procedure (Lim, et al, 2017). There are many possible causes of persistent pain, but one of the most common is a phenomenon referred to as aseptic loosening (Kim, et al, 2014). Aseptic loosening occurs when the bond between the cement layer and either the tibial component of the replacement or tibial bone degrades for reasons other than infection (Lim, et al, 2017). Among the cases of aseptic loosening, it has been shown that the degradation most commonly occurs at the cement-implant interface (Arsoy, et al, 2013). The persistent pain accompanied by this phenomenon is often enough to warrant a revision surgery, and multiple studies have shown that one of the most common reasons for revision surgery is, in fact, aseptic loosening (Postler, et al, 2018), (Sharkey, et al, 2014).

This work refers to the two variations of aseptic loosening as debonding and loosening. More specifically, debonding refers to the variation in which the damage occurs at the interface between the cement and the tibial implant. Loosening refers to the variation in which the damage occurs at the interface between the cement and the tibial bone.

One of the major issues surrounding TKA is the difficulty of diagnosing the cause of persistent pain after the procedure. Currently, the only way to truly diagnose the cause of persistent pain is to operate on the affected knee. This is the major problem we aim to solve. Previous work in our research group sought methods for early detection and diagnosis of aseptic loosening. The work demonstrated the
efficacy of a nondestructive method of structural health monitoring (SHM) known as the electromechanical impedance (EMI) method in detecting loosening as a separate class from the healthy baseline class in simulated TKRs (Ponder, et al, 2018).

SHM is a widely used engineering technique for determining the physical integrity of structures. This is accomplished by comparing measured properties of a structure in an assumed damaged state to those of a known undamaged (baseline) state. Within the field of SHM, one method of acquiring information from a structure is the EMI method. This method utilizes piezoelectric materials to produce high-frequency excitation which allows for the monitoring of a structure’s mechanical impedance response. This is accomplished by monitoring the electrical impedance response of a piezoelectric transducer bonded to the observed structure (Park, et al, 2003). A piezoelectric material is a material that produces a charge proportional to a subjected strain. This property also works in reverse, allowing the piezoelectric material to be subjected to an electric current causing a vibration in the material. This vibration translates from a piezoelectric transducer to the structure it is bonded to. The vibrations of the structure can then be converted into electrical impedance due to the proportionality between the charge and strain of the piezoelectric material (Yan and Chen, 2010). One of the most common piezoelectric materials is lead zirconate titanate (PZT), and this is the material used by DSSL.

The previous work by Ponder et al involved assembling simulated TKRs and artificially damaging, then failing them via both debonding and loosening (Ponder, et al, 2018). Some were left undamaged and used as a baseline case, and in this work, they are referred to as healthy. The healthy TKRs and the TKRs with simulated damage were measured via the EMI method in the damaged state and in the failed state to make up the dataset used in this work. Figure 1 displays a fully constructed simulated TKR consisting of a tibial baseplate cemented to a Sawbones synthetic resected tibia with high-viscosity bone cement, and the approximate position of the PZT sensor is also displayed. More details of the simulated TKR system, experimental setup, and data collection can be found in the work of Ponder et al (Ponder, et al, 2018). In this work, the impedance measurements will be referred to as observations, and the damaged and failed measurements are referred to as weakened and failed, respectively.

The work involving the dataset collected by Ponder was continued by Miller and Anton with the development of classic machine learning models for classifying damage states with varying levels of severity in the simulated TKR system. The goal was to find a combination of model architecture and hyperparameter settings to automate the classification process and eliminate the need for RMSD calculations typically seen with SHM analyses (Miller and Anton, 2021). Five widely used machine learning models were explored and tested including decision trees, k-nearest neighbor (KNN), discriminant analysis (DA), naïve Bayes, and support vector machines (SVM). Leave-one-out cross validation was performed on the trained models to assess their ability to classify the data. It was later determined that the models demonstrated the best performance when trained on the first third of the dataset (i.e. the lower frequencies), with the highest performing model being the KNN (Miller, 2021).

One of the inherent problems with the introduction of machine learning techniques is the black-box nature of many machine learning algorithms. In the previous work, there was no way to truly know how the machine learning algorithms made the classifications. The goal of this work is to expand on the previous work by introducing explainable artificial intelligence (XAI) along with signal processing techniques to develop a more trustworthy model with respect to domain knowledge of the dataset.

The first stage of this work is the selection of a subset of previously studied models for the application of XAI. Once the models are selected, XAI methods are applied to extract...
the features deemed most important for classification of each observation. With the important features identified, knowledge of the dataset as well as knowledge of the machine learning architecture is used to apply pre-processing techniques to the dataset and retrain the selected models on the processed dataset. The final stage of this work is to reapply XAI to the new models for analysis of the most important features.

**Methodology**

**Dataset**

One of the most crucial aspects of this work is understanding the dataset. In the case of the impedance dataset, there are 30 impedance spectrum observations that span a range of 10 kHz to 310 kHz at a resolution of 25 Hz resulting in 12001 impedance datapoints (features) in each observation. Each of the 30 measurements can be classified as one of the following damage states: healthy (7), debonding weakened (6), debonding failed (5), loosening weakened (6), and loosening failed (6). Figure 2 shows an overlay of the healthy impedance observation as well as the various damaged classifications. It is important to note that there are significant variations between the healthy observations and the damaged observations. These variations, or spikes, along the impedance spectrum are visual indicators of damage. Previous work demonstrated that the frequency range of 10 kHz to 110 kHz holds the highest sensitivity to damage, so this study uses only this portion of the dataset resulting in 30 observations, each having 4001 features (Miller, 2021). This portion of the dataset is referred to in this work as lower band dataset.

**Model Selection**

This study utilizes previously trained and tested algorithms for the implementation of XAI methods. For this selection, three previously tested models are chosen spanning all levels of performance, where accuracy is the selected performance metric. This is determined using a leave-one-out cross-validation scheme, where accuracy is the number of correct predictions divided by the number of observations. Previous work training models on the lower band of data and tuning hyperparameters resulted in a KNN model demonstrating the highest accuracy at 76.6%, SVM demonstrated an accuracy that was significantly lower at 63.3%, and DA represents a middle ground model with a demonstrated accuracy of 70% (Miller, 2021). For this reason, KNN, SVM, and DA are the models selected for testing in this study.

**XAI Implementation**

**Selected Technique**

This work uses MATLAB as the primary software for all models and analysis. The XAI method chosen for this feature analysis is the local interpretable model-agnostic explanations (LIME) method. In short, the LIME method determines the most important features of any given observation by training an interpretable model on a new dataset generated by varying the features of the given observation (Ribeiro, et al, 2016). The decision boundaries of the interpretable model correlate to the most important features for the classification of the given observation. LIME was chosen because it is a very well understood method, and it can
provide insight into each individual classification in the dataset. Due to the nonlinear nature of the dataset used in this study, a decision tree was selected as the interpretable model for use with this application of the LIME method because of its similar, nonlinear nature. For this study, the five most important features of every observation are extracted for analysis and comparison with an emphasis on the feature determined to be most important.

**Data Pre-processing**

This work employs pre-processing techniques to adjust the features deemed most important by the LIME method. It is important to note that the machine learning models view each feature independently whereas SHM principles suggest that each feature in an impedance spectrum is dependent on the features surrounding it. This is important because the typical indicators of damage may not necessarily be located at exactly the same feature in every observation. For example, one observation may contain a damage indicator at a frequency of 11000 Hz while another observation contains the same indicator at 11025 Hz. This is common in SHM due to minor, uncontrollable variations in the properties of observed structures. Typical SHM principles recognize that the indicators at these two features are, in essence, the same, but the machine learning used in this study sees the two features as completely independent. For this reason, this study uses averaging as a means to aggregate features and lower the chances of missing relations between damage indicators in the dataset.

The features are aggregated by averaging 1 kHz portions of the lower band dataset. Each aggregated feature is an average of 40 features, and this results in a new dataset consisting of 30 observations, each containing 100 aggregated features. This new dataset is referred to as **aggregated dataset**. Machine learning models are retrained on the aggregated dataset, and the hyperparameters are retuned. The LIME method is again used to extract the most important features of each observation with emphasis on the most important feature. The accuracies of the aggregated models are compared to those of the original models to determine if the feature reduction results in a loss of crucial information from the dataset. The extracted features of the aggregated dataset, again with emphasis on the most important feature, are also compared to those of the original lower band dataset to determine if the selected features make sense with respect to typical SHM domain knowledge.

**Results and Discussion**

**K-Nearest Neighbor**

Discussing first the results of the KNN analysis on the original feature set, the LIME method determined that the most important feature of each observation fell within a range of 13725 Hz – 14700 Hz. Looking at Figure 3, it is immediately noticeable that the feature deemed most important (represented by the black, dashed line) for the classification of the displayed, representative observation is located at a seemingly random location along the impedance spectrum which cannot be explained by typical SHM principles. This is unexpected because, as discussed previously, typical SHM principles would suggest that the indicators of damage in the system correspond with spikes in the plotted impedance spectrum. This observation was selected as representative due to the similar outcome of every observation in the dataset. The representative observation is plotted along with its extracted features. Also plotted are all the healthy observations for visual reference. Note that the same representative observation is selected for displaying all results. The KNN model demonstrated the highest performance with an accuracy of 76.6%, which is interesting considering the model’s most important feature does not correlate with what is considered to be a typical indicator of damage along the impedance spectrum.

Turning next to the results of the KNN analysis on the aggregated dataset, the LIME method determined the most important feature of each observation to be the second aggregated feature which represents a range of 11000 Hz – 11975 Hz. This aggregated feature correlates with the first major spike in the impedance spectrum of each observation. Looking at Figure 4, it is immediately noticeable that the aggregated dataset retained the important spikes in the impedance spectrum. It is also evident that the model is now using an impedance spike as the most important feature to classifying the observation. The same representative observation is displayed for visualization due to the similar results of every observation. The model maintained an accuracy of 76.6% even though the total number of features has been greatly reduced. This may suggest that the dataset has not lost any crucial information with the reduction of features.

![KNN Important Features](image)
Support Vector Machine

Discussing next the results of the SVM analysis, the feature determined to be most important to the classification of each observation was either 11750 Hz or 11775 Hz. Again, this does not correlate with a spike in the impedance spectrum. This is apparent in Figure 5, where it is immediately noticeable that the feature is shifted slightly to the right of the first major spike. The SVM model demonstrates a significantly lower performance than the KNN model with an accuracy of only 63.3%. This low performance makes sense from an engineering standpoint because the model has not successfully identified the significance of the typical indicators of damage for this SHM dataset.

The results of the aggregate SVM study demonstrate that the most important feature identified for each observation is again the second aggregated feature that represents a range of 11000 Hz – 11975 Hz. This feature correlates with the first major impedance spike which can be seen in Figure 6. The aggregated SVM model maintained its performance with an accuracy of 63.3% which is further indication that the aggregating did not result in a significant loss of information from the dataset.

Discriminate Analysis

Finally, turning to the results of the DA analysis, the LIME method determined that the most important feature of each observation fell within a range of 13725 Hz – 13825 Hz. Similar to the results of the KNN study, this again correlates with a seemingly random and unexplainable feature along the impedance spectrum as seen in Figure 7. The DA model demonstrates an accuracy of 70%, which is again interesting considering the feature found to be most important to the model does not appear to hold any SHM significance.

The results of the aggregated DA testing indicate that the most important feature for the classification of each observation was the fourth aggregated feature which represents the frequency range of 14000 Hz – 14975 Hz. This again does not correlate with a typical indicator of damage and instead falls at the same seemingly random feature; the results are displayed in figure 8. The accuracy of the aggregated model did see a decrease in performance with an accuracy of 66.6% which could be attributed to the model again failing to recognize the significance of the impedance spikes. This decrease in accuracy is equivalent to one missed classification when compared to the original model.
Conclusion

The goal of this work was to utilize XAI along with signal processing techniques to develop a more trustworthy classification model for a complex biomedical system. Previously, electromechanical impedance data was collected from a simulated TKR system at varying levels of damage using SHM techniques. Machine learning models were then generated as a means to classify the various damage states present in the dataset. This paper expands that work by implementing XAI as a means to test the ability of machine learning models to accurately identify features indicative of damage. Pre-processing techniques are then used to assist the machine learning models in accurately identifying important features. It has been demonstrated that aggregating features via averaging successfully generated a representative dataset with a reduced number of features and no significant loss of important information. The accuracy of the tested models remained nearly the same before and after the application of aggregation. The important features of the aggregated dataset, as determined using the LIME method, were concluded to be more trustworthy than those of the original dataset. When aggregated, the most important feature of each observation tended to correlate with a spike in the impedance spectrum which is considered to be a typical indicator of damage. In the original lower band dataset, the most important feature did not correlate with a spike in the impedance spectrum and therefore was not located at a typical indicator of damage. In short, XAI successfully added a level of transparency to the models, which resulted in a more trustworthy classification model with respect to domain knowledge of the dataset.

Acknowledgements

The authors would like to acknowledge the efforts of Mr. Robert I. Ponder in collecting the data used for this work, as well as Dr. R. Michael Meneghini in generating the samples from which the data was taken. The authors would also like to acknowledge the efforts of Mr. Brandon A. Miller in testing the machine learning models used for this work.

References


Ribeiro, M. T. et al., 2016, "Why should I Trust You?" Explaining the Predictions of any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp.1135-1144).

Sharkey, P. F., MD et al., 2014. Why are Total Knee Arthroplasties Failing Today—Has Anything Changed After 10 Years? The Journal of Arthroplasty vol. 29, (9): 1774-1778.