# Determinantal Point Processes (DPPs) Improve KNN Classifier Performance in Bioimpedance Analysis

**Mali Halac**

## 1 Introduction

Electrical impedance measurements have emerged as a potent tool in the differentiation and analysis of physiological tissues. The unique impedance characteristics of various tissues offer a non-invasive approach to study their properties and conditions. This paper focuses on leveraging these impedance differences to distinguish between nerve and muscle tissues, a distinction that is vital in numerous medical and research applications.

The central challenge in tissue impedance analysis lies in the complexity and variability of the data, exacerbated by intersubject differences. Traditional classification methods often struggle with generalizability across different subjects due to these inherent variabilities. Therefore, there is a pressing need for robust classification techniques that can effectively handle this diversity and offer accurate, generalizable results.

In recent years, machine learning has provided a pathway to address these challenges. Among various algorithms, the K-Nearest Neighbor (KNN) classifier stands out for its simplicity and effectiveness in classification tasks. However, the performance of KNN is highly dependent on the features used for classification. This brings us to the critical aspect of feature selection, where the goal is to identify features that are most informative and discriminative for the task at hand.

This paper applies Determinantal Point Processes (DPPs) to feature selection in the context of tissue impedance analysis. DPPs are probabilistic models that focus on selecting a diverse set of features, thereby reducing redundancy and enhancing the classifier's ability to generalize across different subjects [1,2]. Coupled with a Radial Basis Function (RBF) kernel, DPPs provide a principled way to select the most informative and diverse features from impedance data.

Our study demonstrates how the integration of DPPs with the KNN classifier leads to significant improvements in the accuracy and generalizability of tissue type classification. This approach not only provides a more robust method for analyzing tissue impedance data but also opens new avenues for applying DPPs in biomedical data analysis.

The subsequent sections detail the methodology adopted, the implementation of DPPs using the DPPy library in Python, the application of the RBF kernel, and the configuration of the KNN classifier. We then present our results, showcasing the efficacy of this combined approach, followed by a discussion on the implications of our findings.

## 2 Background

### 2.1 Determinantal Point Processes (DPPs)

Determinantal Point Processes (DPPs) are probabilistic models that quantify the likelihood of selecting a subset from a set of items, emphasizing diversity. The probability of observing a subset $Y \subseteq \mathcal{X}$ in a DPP is proportional to the determinant of a positive semidefinite kernel matrix $K$:

$$P(Y \subseteq \mathcal{X}) \propto \det(K_Y) \tag{1}$$

where $K_Y$ is the matrix derived by selecting rows and columns from $K$ indexed by the elements of $Y$. More information could be found in Kulesza and Taskar (2012), Kulesza and Taskar (2011), and Edelman (In progress). In this study, we used the DPPy library [4] in Python to implement DPPs, particularly for selecting frequency points in impedance data.

### 2.2 Radial Basis Function (RBF)

The Radial Basis Function (RBF) is a widely-used kernel function in machine learning, defined as:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{2}$$

where $x$ and $x'$ represent two points in the input space, $\sigma$ is the bandwidth parameter, and $\|\cdot\|^2$ denotes the Euclidean distance squared. The RBF kernel was utilized in our DPP model to assess the similarity between different frequency points.

### 2.3 K-Nearest Neighbor (KNN) Classifier

The K-Nearest Neighbor (KNN) classifier is a straightforward yet effective algorithm for classification tasks. It classifies a test point $x$ by identifying the $k$ nearest points in the training set and assigning $x$ to the majority class among these points:

$$f(x) = \arg\max_c \sum_{i=1}^{k} \mathbb{I}(c_i = c) \tag{3}$$

where $\mathbb{I}$ is the indicator function, and $c_i$ is the class label of the $i$-th nearest neighbor of $x$.

## 3 Methods & Results

In this study, we conducted a frequency sweep ranging from 1 kHz to 100 kHz, using 1024 frequency points. This process generated frequency-specific impedance curves, as depicted in Fig. 1. Our initial hypothesis was that the electrical impedance characteristics of physiological tissues could provide insights into specific tissue types. To explore this, we employed a K-Nearest Neighbor (KNN) classifier to distinguish between impedance measurements derived from nerve and muscle tissues.

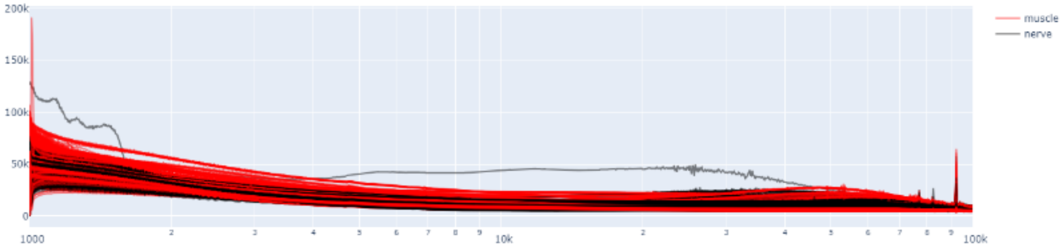
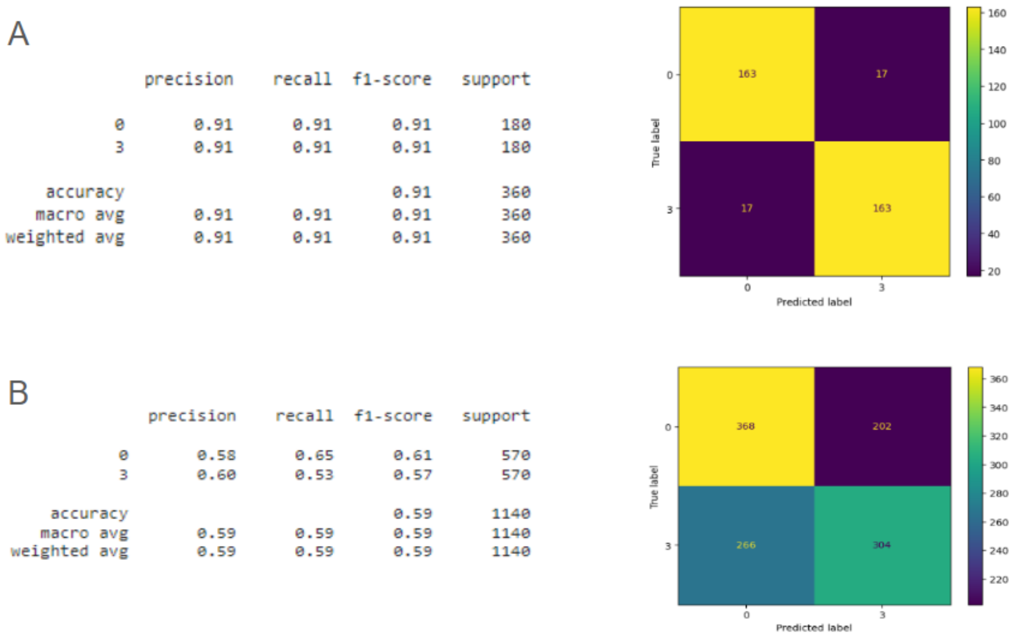


Figure 1: **Frequency sweep of rat nerve and muscle tissues in range [1,100] kHz.**

Our preliminary analysis involved impedance measurements from a single rat, encompassing the sciatic nerve, facial nerve, and adjacent muscles on both sides. We allocated the data into training and testing sets, with a split of 70% and 30%, respectively. Both the training and testing sets were z-score normalized. The initial application of the KNN classifier to this dataset yielded a promising accuracy of 91%, as shown in Fig. 3A. To validate these findings, we extended the testing to include impedance measurements from three additional rats. However, upon applying the same classifier to this expanded dataset, we noted a significant reduction in accuracy, which fell to 59%. This suggests that while the classifier performs well with data from a single subject, its effectiveness diminishes when generalized across multiple subjects.



```
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       180
           3       0.91      0.91      0.91       180

    accuracy                           0.91       360
   macro avg       0.91      0.91      0.91       360
weighted avg       0.91      0.91      0.91       360
```

```
              precision    recall  f1-score   support

           0       0.58      0.65      0.61       570
           3       0.60      0.53      0.57       570

    accuracy                           0.59      1140
   macro avg       0.59      0.59      0.59      1140
weighted avg       0.59      0.59      0.59      1140
```

Figure 2: **K-Nearest Neighbor (KNN) classifier trained on one rat. A. Tested on the same rat. B. Tested on new rats.**

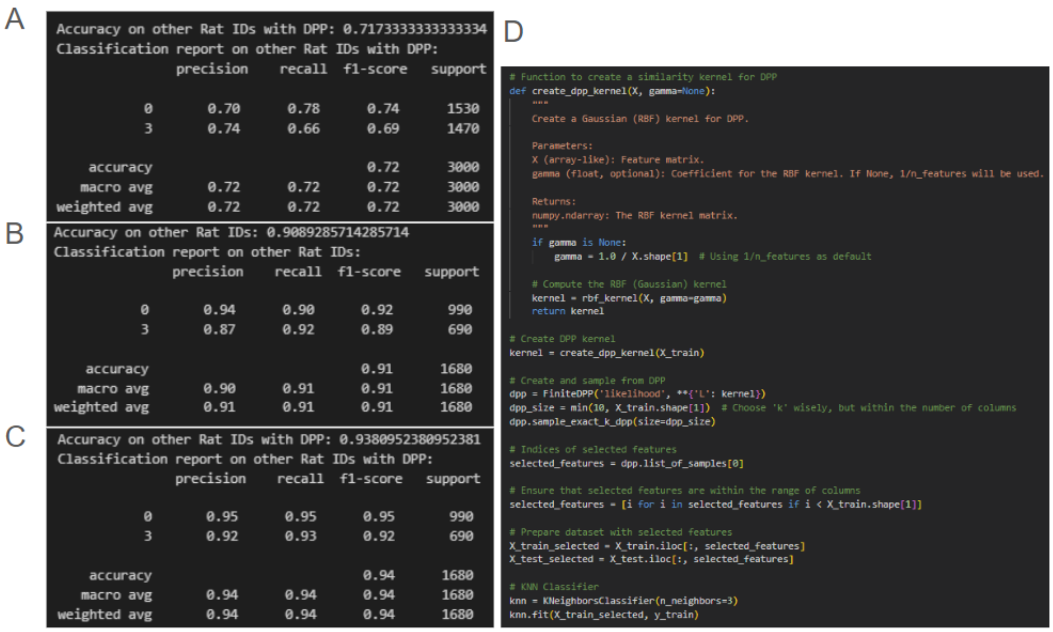
Building on our initial findings, we explored the potential of utilizing a determinantal point process (DPP) model to enhance the intersubject performance of our KNN classifier. It is pertinent to recall here that DPPs are known for modeling the inherent repulsion within datasets. In the context of our study, applying a DPP model to the frequency sweep of impedance was hypothesized to identify the most diverse frequency points, which are likely to be robust against variations among different subjects.

Acting on this hypothesis, we employed a sampling DPP with a radial basis function kernel for the purpose of feature selection, limiting the selection to ten features (n=10). Subsequently, we retrained the KNN classifier using this refined set of features. When this updated classifier was tested on the impedance measurements from the three new rats, we observed a marked improvement in performance. As illustrated in Fig. 3A, the accuracy of the classifier increased to 72%.

The initial improvement in accuracy, while encouraging, was still not optimal. To address this, we conducted a meticulous examination of our training data to identify and exclude any potential outliers or mislabeled entries. During this scrutiny, we identified certain instances where the measurements were compromised. Notably, some readings were affected by the electrodes not making proper contact with the tissue. This was sometimes due to air gaps or surgical liquids forming barriers, which significantly altered the impedance curves. Other outliers included measurements inadvertently taken in air when the operator was unprepared, instead of the intended tissue impedance.

Furthermore, we noticed that the equipment used for measuring electrical impedances was susceptible to noise at lower frequencies. To mitigate this, we implemented a high-pass filter, excluding frequencies below 5,500 Hz. This adjustment reduced the total frequency points from 1024 to 688.

After these modifications and the removal of outliers and mislabeled data, we retrained the KNN classifier using data from one rat and tested it on three new rats. This led to a substantial increase in accuracy, as shown in Fig. 3B, with a jump from 72% to 91%. Further enhancement was achieved when we reapplied the DPP sampling algorithm, using a radial basis function as the kernel. This resulted in an additional 3% improvement in accuracy, as illustrated in Fig. 3C. It's important to highlight that the DPP sampling algorithm, with merely 10 selected features, outperformed the earlier model that used 688 features. This underscores the efficiency of the DPP algorithm in eliminating redundant features that do not significantly contribute to tissue discrimination.



```
Accuracy on other Rat IDs with DPP: 0.7173333333333334
Classification report on other Rat IDs with DPP:
              precision    recall  f1-score   support

           0       0.70      0.78      0.74      1530
           3       0.74      0.66      0.69      1470

    accuracy                           0.72      3000
   macro avg       0.72      0.72      0.72      3000
weighted avg       0.72      0.72      0.72      3000
```

```
Accuracy on other Rat IDs: 0.9089285714285714
Classification report on other Rat IDs:
              precision    recall  f1-score   support

           0       0.94      0.90      0.92       990
           3       0.87      0.92      0.89       690

    accuracy                           0.91      1680
   macro avg       0.90      0.91      0.91      1680
weighted avg       0.91      0.91      0.91      1680
```

```
Accuracy on other Rat IDs with DPP: 0.9380952380952381
Classification report on other Rat IDs with DPP:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95       990
           3       0.92      0.93      0.92       690

    accuracy                           0.94      1680
   macro avg       0.94      0.94      0.94      1680
weighted avg       0.94      0.94      0.94      1680
```

```python
# Function to create a similarity kernel for DPP
def create_dpp_kernel(X, gamma=None):
    """
    Create a Gaussian (RBF) kernel for DPP.

    Parameters:
    X (array-like): Feature matrix.
    gamma (float, optional): Coefficient for the RBF kernel. If None, 1/n_features will be used.

    Returns:
    numpy.ndarray: The RBF kernel matrix.
    """
    if gamma is None:
        gamma = 1.0 / X.shape[1]  # Using 1/n_features as default

    # Compute the RBF (Gaussian) kernel
    kernel = rbf_kernel(X, gamma=gamma)
    return kernel

# Create DPP kernel
kernel = create_dpp_kernel(X_train)

# Create and sample from DPP
dpp = FiniteDPP('likelihood', **{'L': kernel})
dpp_size = min(10, X_train.shape[1])  # Choose 'k' wisely, but within the number of columns
dpp.sample_exact_k_dpp(size=dpp_size)

# Indices of selected features
selected_features = dpp.list_of_samples[0]

# Ensure that selected features are within the range of columns
selected_features = [i for i in selected_features if i < X_train.shape[1]]

# Prepare dataset with selected features
X_train_selected = X_train.iloc[:, selected_features]
X_test_selected = X_test.iloc[:, selected_features]

# KNN Classifier
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train_selected, y_train)
```

Figure 3: **A. Determinantal point processes improve intersubject performance B. KNN classifier performance after removing mislabels and outliers. B. KNN performance with DPP sampling after removing mislabels and outliers. D. Base code used for DPP feature selection and KNN classifier.**

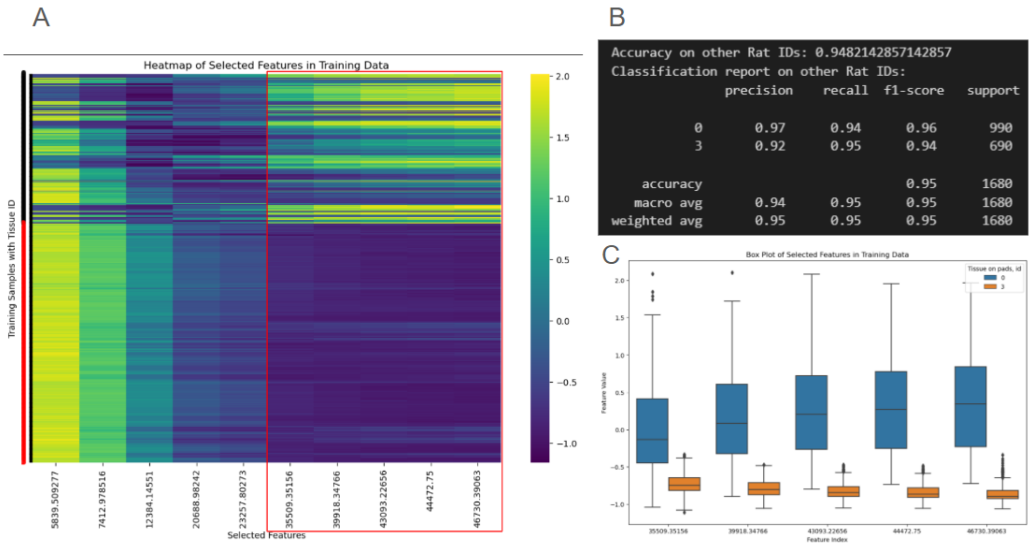
Our next focus was on the ten features selected by the DPP sampling algorithm, which contributed to the enhanced performance of the KNN classifier. As depicted in Fig. 4A, these selected features are represented in a heatmap format. Each row in the heatmap corresponds to a unique measurement, with red and black colors on the left denoting muscle and nerve tissues, respectively. The specific frequencies of these selected features, listed at the bottom of each column, are 5.8 kHz, 7.4 kHz, 12.3 kHz, 20.6 kHz, 23.2 kHz, 35.5 kHz, 39.9 kHz, 43.0 kHz, 44.4 kHz, and 46.7 kHz. Notably, the five frequencies (n=5) that offer the highest class separation are distinctly marked with a red rectangle on the heatmap. This visualization clearly illustrates that these frequencies contribute significantly to class differentiation. Intrigued by this observation, we further narrowed our focus to these five frequencies, as highlighted in Fig. 4A, and questioned the outcome of using them as the sole features for the DPP sampling (n=5). Fig. 4B showcases the classifier's performance when trained exclusively on these five highly discriminatory features. Remarkably, this resulted in an additional 1% improvement in accuracy, reinforcing our initial hypothesis that DPP effectively eliminates redundant features, which otherwise act as noise, thus sharpening the classifier's focus on class-discriminatory features.

Following this performance enhancement, we delved into whether these five frequency points identified by the DPP had an amplitude threshold that distinguished between nerve and muscle tissues. Fig. 4C presents the distribution of impedance amplitudes for both tissue types at the selected frequencies. This distribution confirms that the DPP algorithm selectively identified frequency points exhibiting high class separation.

## 4 Discussion

Our study has demonstrated the significant impact of Determinantal Point Processes (DPPs) in improving the generalization capabilities of classifiers to new subjects. This finding is crucial, especially in the context of physiological data analysis, where individual variability can often pose a challenge to the robustness and accuracy of classification algorithms.

One of the key strengths of DPPs, as evidenced in our research, is their ability to effectively eliminate redundancy in the feature set. By focusing on the most diverse and informative features, DPPs



```
Accuracy on other Rat IDs: 0.9482142857142857
Classification report on other Rat IDs:
              precision    recall  f1-score   support

           0       0.97      0.94      0.96       990
           3       0.92      0.95      0.94       690

    accuracy                           0.95      1680
   macro avg       0.94      0.95      0.95      1680
weighted avg       0.95      0.95      0.95      1680
```

Figure 4: **DPPs select the most diverse features.** A. Heatmap of selected features. Each row correspond to a different measurement. Each column represents the DPP selected feature. The specific frequency (Hz) of the selected features can be seen at the bottom of each column. The red bard on the left indicates the measurements that correspond to a muscle tissue. The black bar on the left indicates the measurements corresponding to a nerve tissue. B. The performance of KNN classifier after being trained on DPP selected features (n=5) C. DPP selected features have an amplitude threshold. 0 in the legend corresponds to nerve tissues, 3 corresponds to muscle tissues.

streamline the classifier's input, reducing the noise and irrelevant information that can often lead to overfitting. This reduction of redundancy not only simplifies the model but also enhances its interpretability and efficiency.

Furthermore, the selection of the most diverse features by DPPs is a cornerstone in their contribution to classifier improvement. By prioritizing features that provide the highest class separation, DPPs ensure that the classifier's training is focused on the most relevant and distinctive aspects of the data. This approach is particularly advantageous in distinguishing between subtle variations in complex datasets, such as those encountered in differentiating between tissue types in our study.

Overall, the application of DPPs in our research has underscored their potential as a powerful tool for feature selection in machine learning. Their ability to enhance generalization, reduce redundancy, and select diverse features makes them an invaluable asset in developing more accurate, robust, and efficient classifiers.

## References

[1] Kulesza, A., & Taskar, B. (2012). Determinantal Point Processes for Machine Learning. In Foundations and trends in machine learning (Vol. 16, Issues 2-3, pp. 123–286). now. https://doi.org/10.1561/2200000044

[2] Kulesza, A., & Taskar, B. (2011). k-DPPs: Fixed-Size Determinantal Point Processes. Proceedings of the 28th International Conference on Machine Learning (ICML).

[3] Alan Edelman. (In progress). Random Matrix Theory.

[4] Gautier, G., Polito, G., Bardenet, R., & Valko, M. (2019). DPPy: DPP Sampling with Python. Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS).