INTRODUCTION

In this research, machine learning methodologies were employed to investigate the correlation between the design parameters of an antenna and its electromagnetic performance. The analysis focused on the S11 parameter, which signifies signal loss and is pivotal for antenna efficiency, across various frequency points using dimensionality reduction and regression modeling techniques. Initially, the real and imaginary components of the S11 parameter, along with the design parameters, were extracted from CSV files, followed by the computation of basic statistics and a correlation matrix for the input dataset. To facilitate data reduction and identify the components contributing the most variance, Principal Component Analysis (PCA) was applied after standardizing the input data. A cumulative explained variance graph was generated to depict the number of components accounting for 90% of the total variance.

Subsequently, the magnitude of S11 was derived from its real and imaginary parts, with the minimum S11 magnitude and the associated frequency identified for each design iteration. Furthermore, the antenna performance was illustrated through plots of the S11 magnitude against frequency indices for both the initial six designs and a selected group of fifteen designs. The dataset that had been reduced via PCA served as input features in the regression model aimed at estimating S11 at designated frequency points. Consequently, a linear regression model was developed to predict the real component of S11 at specified frequencies, with model efficacy evaluated by juxtaposing the estimated S11 values against the actual measurements. The findings of this research underscore the efficacy of PCA and regression models in elucidating the influence of antenna design parameters on electromagnetic performance and present a significant data-driven methodology for forecasting antenna performance.

What is S11?

S-parameters are fundamentally characterized as a function of frequency. Specifically, S11 quantifies the amount of power reflected from the antenna, and is commonly referred to as the reflection coefficient, sometimes denoted as gamma, or return loss. When S11 is equal to 0 dB, it indicates that all incident power is reflected by the antenna, with no power being effectively radiated.

What is PCA?

Principal Component Analysis (PCA) is a prominent technique employed in both dimensionality reduction and machine learning. It aims to condense a large dataset into a more manageable form while preserving essential patterns and trends within the data. This process of reducing variables inevitably leads to a trade-off between accuracy and simplicity; however, the key to successful dimensionality reduction lies in sacrificing a minor degree of accuracy to achieve improved clarity. Smaller datasets facilitate exploration and visualization, thereby enabling machine learning algorithms to analyze data points more efficiently without the complication of extraneous variables.

What is Linear Regression? Why is Linear Regression important?

Linear regression is a widely recognized supervised machine learning algorithm that establishes a linear relationship between a dependent variable and one or more independent variables by fitting an equation to observed data. One of the most significant advantages of linear regression is its interpretability. The resulting equation provides distinct coefficients that clarify the influence of each independent variable on the dependent variable, fostering a comprehensive understanding of the underlying dynamics. Moreover, its simplicity is a considerable asset; linear regression is transparent, straightforward to implement, and serves as a foundational concept for understanding more sophisticated algorithms.

Data Generation

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
```

```
#Load Data
input_data = pd.read_csv('hw1_input.csv')
real_data = pd.read_csv('hw1_real.csv')
image_data = pd.read_csv('hw1_img.csv')

#Display the first few rows to verify data
print("Real Data (S11 Real Part):")
print(real_data.head())
print("\nInput Data (Geometric Parameters):")
print(input_data.head())
print("\nImage Data (S11 Imaginary Part):")
print(image_data.head())

#Data Summary and Correlation Matrix
summary = input_data.describe()
print("\nData Summary:")
print(summary)

correlation_matrix = input_data.corr()
print("\nCorrelation Matrix:")
print(correlation_matrix)

#Standardize Data for PCA
scaler = StandardScaler()
scaled_input = scaler.fit_transform(input_data)

#Applying PCA and Cumulative Explained Variance
pca = PCA()
pca_components = pca.fit_transform(scaled_input)
explained_variance = pca.explained_variance_ratio_
cumulative_variance = np.cumsum(explained_variance)

#Print explained variance for each component and cumulative explained variance
print("\nExplained Variance by Each Component:")
```

```
print(explained_variance)
print("\nCumulative Explained Variance:")
print(cumulative_variance)

#Plot Cumulative Explained Variance
plt.figure(figsize=(10, 6))
plt.plot(cumulative_variance, marker='o')
plt.axhline(0.90, color="r", linestyle="--", label="90% Explained Variance")
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
plt.title('PCA - Cumulative Explained Variance by Number of Components')
plt.legend()
plt.grid(True)
plt.show()

#Determine number of components explaining 90% variance
pca_90 = PCA(n_components=0.90)
pca_90_components = pca_90.fit_transform(scaled_input)
print(f"Number of components to explain 90% variance: {pca_90.n_components_}")

#S11 Magnitude Calculation and Min Magnitude by Design
s11_magnitude = np.sqrt(real_data**2 + image_data**2)
min_indices = s11_magnitude.idxmin(axis=1)
min_values = s11_magnitude.min(axis=1)

#Display min S11 magnitude and corresponding frequency index for each design
result = pd.DataFrame({
    'Design': s11_magnitude.index,
    'Min Frequency Index': min_indices,
    'Min S11 Magnitude': min_values
})
print("\nMinimum S11 Magnitude by Design:")
print(result)
```

```
#Plot S11 Magnitude Across Frequency for Selected Designs
num_designs_to_plot = 15
for i in range(num_designs_to_plot):
    s11_values = s11_magnitude.iloc[i, :]
    plt.plot(s11_values, label=f"Design {i+1}")
plt.xlabel("Frequency Index")
plt.ylabel("S11 Magnitude")
plt.title("S11 Magnitude Across Frequency Indices for Selected Designs")
plt.legend(loc="best")
plt.grid(True)
plt.show()

# PCA for Regression Task
selected_frequencies = [70, 120]  # Select specific frequency indices for regression target
y = real_data.iloc[:, selected_frequencies]  # Real part of S11 at selected frequencies

#Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(pca_90_components, y, test_size=0.2,
random_state=42)

#Linear Regression Model
model = LinearRegression()
model.fit(X_train, y_train)

#Predictions on test set and model evaluation
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"\nLinear Regression Model Performance:")
print(f"Mean Squared Error (MSE): {mse}")
print(f"R^2 Score: {r2}")

#Visualizing Predictions vs. True Values
```

```
for i, freq in enumerate(selected_frequencies):
    plt.figure(figsize=(8, 5))
    plt.scatter(y_test.iloc[:, i], y_pred[:, i], alpha=0.7, label=f'Frequency {freq}')
    plt.plot([y_test.iloc[:, i].min(), y_test.iloc[:, i].max()],
             [y_test.iloc[:, i].min(), y_test.iloc[:, i].max()], 'r--', label="Ideal Fit")
    plt.xlabel("True S11 (Real Part)")
    plt.ylabel("Predicted S11 (Real Part)")
    plt.title(f"S11 Prediction Performance - Frequency {freq}")
    plt.legend()
    plt.grid(True)
    plt.show()
```

Results and Analysis

Data Summary:

	length of patch	dielectric constant of solder resist layer
count	385	385
mean	3.569210	3.521911
std	0.966173	0.871233
Min	1.805658	2.001679
25%	2.755534	2.783710
50%	3.637716	3.480916
75%	4.369311	4.278575
Max	5.199919	4.999950

We use statistics to gain insight into the variation and distribution of parameters, which allows us to acquire more information about the variables in the dataset. For example, the minimum value of the "length of patch" parameter is observed as 1.805, while the maximum value is 5.200. These values are useful for determining the boundary limits of the antenna design. The mean represents the center of these values, and the standard deviation provides information about the extent of the distribution of these variables. When examining the standard deviation given for two variables, we observe that the "length of patch" has a wider

spread of data. Looking at the interquartile range, we can also infer that our data is not concentrated in extreme values but rather distributed across the range of values.

Based on this information, we can suggest that variables with high variability should be optimized. By identifying which variables are critical for us, we can more easily improve model performance.

Explained Variance by Each Component & Cumulative Explained Varaince:

	Explained Variance	Cumulative Explained Variance
Component1	0.20715348	0.20715348
Component2	0.11070825	0.31786173
Component3	0.10003384	0.41789557
Component4	0.09607237	0.51396794
Component5	0.09187079	0.60583873
Component6	0.08845826	0.69429699
Component7	0.08565904	0.77995603
Component8	0.08165053	0.86160656
Component9	0.07339863	0.93500519
Component10	0.05830268	0.99330787
Component11	0.00669213	1.00

When we look at the results of ‘Explained Variace by Each Component’ and ‘Cumulative Explained Variance’, we observe that we can achieve approximately 90% success using the first 9 components. We can say that the first component has the highest assumption value. In addition, the last component has the least effect. We can clarify 51% of the data with the first four components. Looking at the numerical results, we can say that the slope gradually decreases during the graph drawing since we know that the graph increases rapidly at the beginning and the increase decreases as the number of components increases. Also, we know that any cumulative value will end at point 1. When PCA is applied generally, it is found sufficient to clarify 90% or 95% of the data because it is a sufficient rate to understand and model the data.

Number of components to explain 90% variance: 9

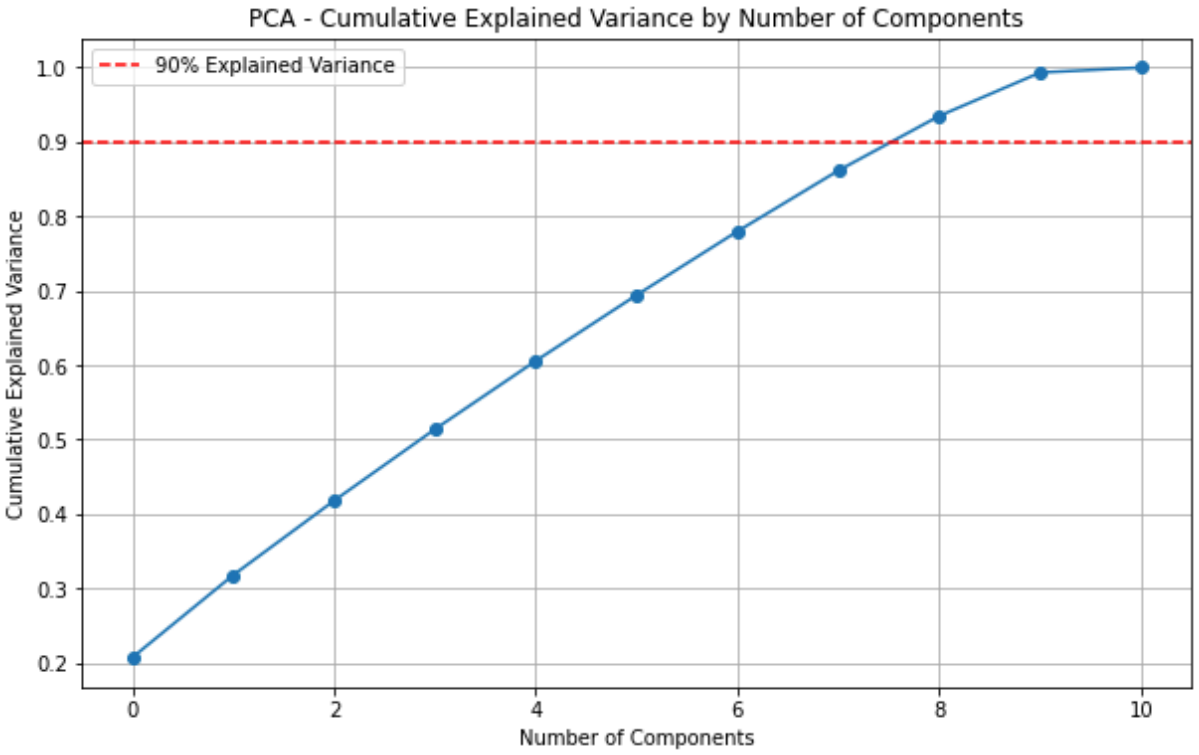
Linear Regression Model Performance:

Mean Squared Error (MSE)	0.11040382414865232
R_2 Score	0.7459607716190384

The Mean Squared Error (MSE) provides an indication of the model's error level, while the R_2 value signifies the model's accuracy rate. A notable observation is that the MSE increases when the number of components is reduced from ten to nine. (This phenomenon may be readily verified through code adjustments that alter the number of components. The accompanying graph below elucidates this relationship, allowing for a more comprehensive understanding of how component variation affects the accuracy rate.)

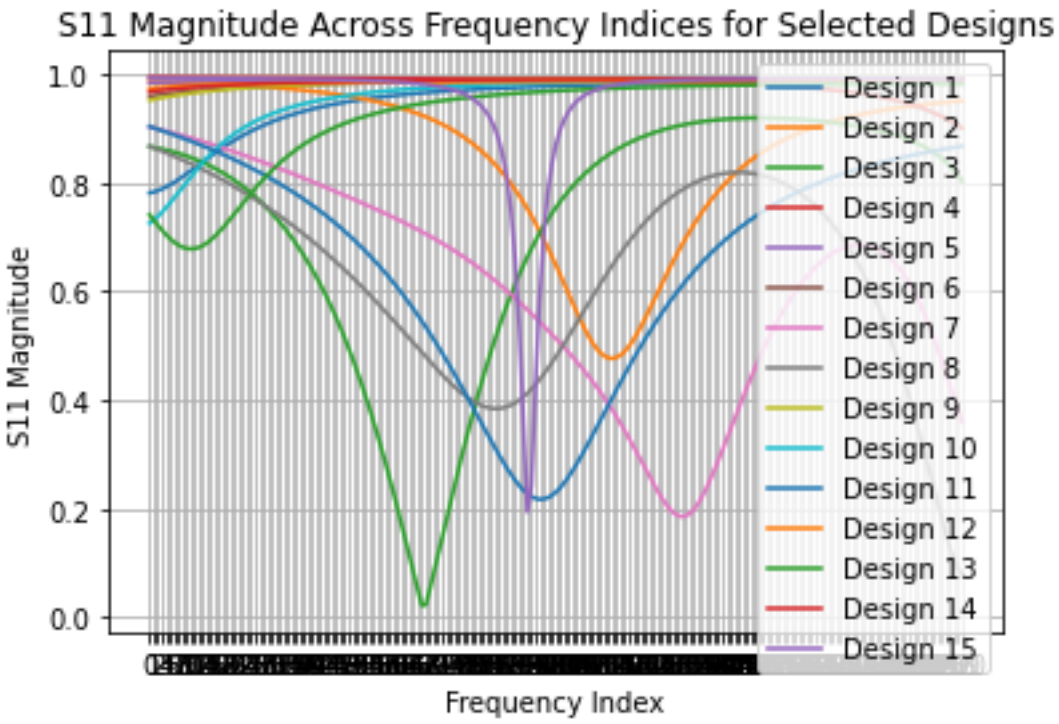
An increased MSE indicates a wider error margin in the model's estimations. Concurrently, an examination of the R_2 value reveals a corresponding decline, suggesting that accuracy deteriorates as the number of components diminishes. This correlation leads to the inference that a reduction in component quantity adversely impacts the model's estimation capabilities. Therefore, it is advisable to retain the ten-component configuration to ensure the accuracy of the S11 response in antenna design.

Nonetheless, if the model does not yield satisfactory estimations, it can be evaluated through adjustments of plus or minus five percent. A deeper statistical analysis can further elucidate the error margin. Furthermore, the significance of reducing the dimensional component count is paramount; failure to do so may hinder the model's capacity for making precise predictions.

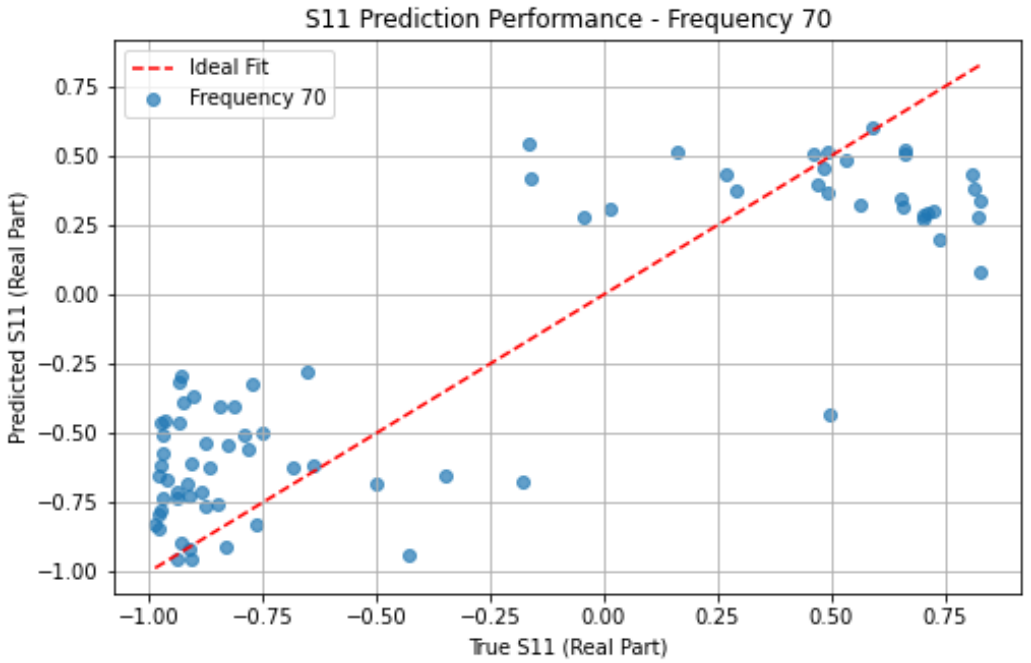


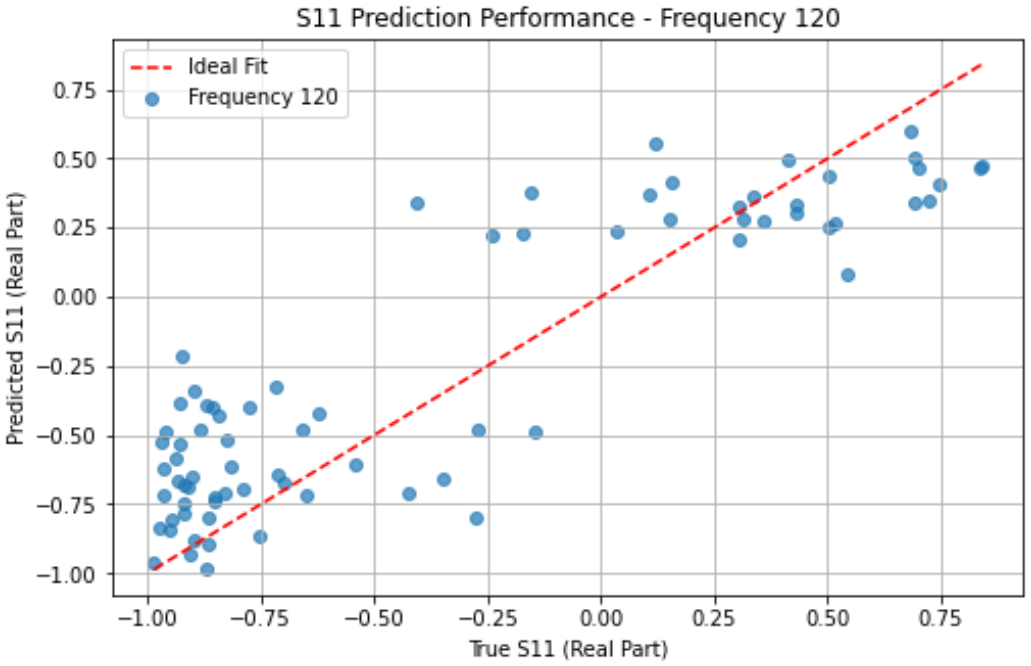
Upon examining the graph, it becomes evident that the cumulative explained variance exhibits a rapid increase with the initial variables. This observation leads us to conclude that the early values exert a significant influence on the model. However, after reaching a specific threshold, the progression of the graph tends toward linearity, indicating that the impact of subsequent components is relatively low in the context of the model's estimation. Notably, we determine that an optimal efficiency from our model is achievable with ten variables. Additionally, it is crucial to acknowledge that the variance values represented are cumulative. Analyzing the cumulative variance as each component is added provides clarity regarding the diminishing effect of variance after a certain number of components. Specifically, at nine components, the model demonstrates an ability to predict with an accuracy of 90%. The intersection of the red line with the blue graph illustrates the prediction accuracy corresponding to the number of components utilized. It is important to note a discrepancy observed in the graph; this can be attributed to the generalized approach employed in the graph's construction, contrasting with the more nuanced method used during the coding process. Consequently, visual representations may not consistently reflect the most accurate results. Furthermore, it can be posited that the model delineated by the blue graph achieves its maximum accuracy after a particular count of components. Such graphical representations

provide a more tangible understanding than a mere numerical context. Is not life all about the bridges we build between the abstract and the concrete?



Interpretation of the image: The S11 magnitude approaches approximately 0, indicating that the antenna effectively propagates the signal at its designated frequency while operating with minimal loss. An S11 value near 0 denotes an optimal frequency range for performance. The analysis of the image reveals a broad frequency spectrum, facilitating the identification of points at which efficiency increases or decreases. Given that the minimum S11 value correlates with peak antenna performance, the insights derived from this wide frequency range are particularly significant for evaluating antenna efficiency.





The graphs we obtained illustrate the overlap between the estimated S11 values and the actual S11 values. However, when working with a complex dataset, we find that using the linear regression method may not lead to a successful predictive model. The red line represents the ideal fit line, while the blue dots along the line correspond to the data points that match our model's actual data. To improve our predictions, we can explore alternative methods, such as adding polynomial features and non-linear regression models, which can help align the model more closely with the real data.

Conclusion

In this project, dimensionality reduction and regression techniques were applied to model the relationship between antenna design parameters and electromagnetic performance (S11 parameter). Thanks to the amount of data, the PCA technique was used to achieve a lower-cost and more time-efficient model. Recognizing that predicting the S11 parameter was our critical goal, we determined through the PCA method that 9 components would provide us with 90% variance, representing an acceptable level of accuracy. Then, a linear regression model was applied to make predictions.

The PCA modeling allowed us to filter out data with similar characteristics and work with data that had a high impact on the performance. Specific frequency points were selected to measure the accuracy of our linear regression model. However, upon analyzing the results, we found that the linear regression model was not successful. This outcome suggests that when working with complex datasets or data that cannot be explained by linear equations, it is necessary to proceed with non-linear models. Additionally, using Kernel PCA instead of the standard PCA method could better capture relationships in complex data structures. This project demonstrates that evaluating the performance of models is more important than merely applying them.

References

1. Built In. (n.d.). *A step-by-step explanation of principal component analysis*. Built In. Retrieved November 13, 2024, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

2. GeeksforGeeks. (n.d.). *ML | Linear regression*. GeeksforGeeks. Retrieved November 13, 2024, from https://www.geeksforgeeks.org/ml-linear-regression/

3. Machine Learning Türkiye. (2019, January 18). *Principal component analysis (PCA)*. Medium. Retrieved November 13, 2024, from https://medium.com/machine-learning-t%C3%BCrkiye/pca-b4d745de33f3

4. Antenna Theory. (n.d.). *S-parameters and antenna theory*. Antenna Theory. Retrieved November 13, 2024, from https://www.antenna-theory.com/definitions/sparameters.php

5. OpenAI. (2024). ChatGPT (Version 4). OpenAI. https://chat.openai.com/