Progress - Till Date

Peptides:

Peptides refer to short chains of amino acids, the building blocks of proteins. The study used a dataset consisting of 10,075 allergens and an equal number of non-allergens. These allergens and non-allergens are likely to be proteins, and peptides are fragments of these proteins.

IgE Epitopes:

The study involves the use of 10,451 IgE epitopes to identify antigenic regions in proteins. IgE epitopes are specific regions within proteins that are recognized by the immune system, particularly the IgE antibodies associated with allergic reactions.

Dataset Size for Model Training:

1. Allergens: 10,075
2. Non-allergens: 10,075
3. IgE epitopes: 10,451

Results of the ML Model - Before Extracting the five E-descriptors

1. **Random Forest - 95%**

```
Overall Accuracy: 95.40% Using Random Forest

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.96      0.95      1338
           1       0.96      0.95      0.95      1317

    accuracy                           0.95      2655
   macro avg       0.95      0.95      0.95      2655
weighted avg       0.95      0.95      0.95      2655
```

2. **LogisticRegression - 81%**

```
Overall Accuracy: 81.43% using LogisticRegression Model

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.84      0.82      1338
           1       0.83      0.79      0.81      1317

    accuracy                           0.81      2655
   macro avg       0.81      0.81      0.81      2655
weighted avg       0.81      0.81      0.81      2655
```

Results of the DL Model - Before Extracting the five E-descriptors

1. **Sequential Neural Network Algorithm - 90 %**

```
Validation Accuracy: 90.36% Using Sequential Neural Network Algorithm

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91      1338
           1       0.92      0.89      0.90      1317

    accuracy                           0.90      2655
   macro avg       0.90      0.90      0.90      2655
weighted avg       0.90      0.90      0.90      2655
```

2. **Hyper-Parameter Tuning Algorithm - 91%**

```
Epoch 1/10
299/299 [==============================] - 0s 716us/step - loss: 0.4191 - accuracy: 0.8171 - val_loss: 0.3339 - val_accuracy: 0.8738
Epoch 2/10
299/299 [==============================] - 0s 477us/step - loss: 0.3067 - accuracy: 0.8788 - val_loss: 0.2960 - val_accuracy: 0.8851
Epoch 3/10
299/299 [==============================] - 0s 501us/step - loss: 0.2751 - accuracy: 0.8928 - val_loss: 0.2708 - val_accuracy: 0.8974
Epoch 4/10
299/299 [==============================] - 0s 489us/step - loss: 0.2533 - accuracy: 0.9030 - val_loss: 0.2530 - val_accuracy: 0.9021
Epoch 5/10
299/299 [==============================] - 0s 489us/step - loss: 0.2382 - accuracy: 0.9093 - val_loss: 0.2539 - val_accuracy: 0.8974
Epoch 6/10
299/299 [==============================] - 0s 488us/step - loss: 0.2249 - accuracy: 0.9142 - val_loss: 0.2353 - val_accuracy: 0.9134
Epoch 7/10
299/299 [==============================] - 0s 484us/step - loss: 0.2130 - accuracy: 0.9197 - val_loss: 0.2307 - val_accuracy: 0.9115
Epoch 8/10
299/299 [==============================] - 0s 482us/step - loss: 0.2047 - accuracy: 0.9217 - val_loss: 0.2249 - val_accuracy: 0.9143
Epoch 9/10
299/299 [==============================] - 0s 478us/step - loss: 0.1950 - accuracy: 0.9254 - val_loss: 0.2197 - val_accuracy: 0.9181
Epoch 10/10
299/299 [==============================] - 0s 472us/step - loss: 0.1857 - accuracy: 0.9295 - val_loss: 0.2194 - val_accuracy: 0.9134
83/83 [==============================] - 0s 262us/step
Overall Accuracy: 91.00%

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.91      0.91      1338
           1       0.91      0.91      0.91      1317

    accuracy                           0.91      2655
   macro avg       0.91      0.91      0.91      2655
weighted avg       0.91      0.91      0.91      2655
```

Discussion with Sir

During a recent discussion with Sir, we delved into a comparative analysis between our current group work and our previous endeavor. The focus was primarily on examining the efficacy of various datasets and sub-datasets, as well as the incorporation of libraries for feature engineering.

Evaluation of Datasets:

- Discussed the datasets utilized in both projects, highlighting their respective strengths and weaknesses.
- Compared the relevance and suitability of data sources in addressing project objectives

Assessment of Sub-datasets:

- Analyzed the segmentation of data into sub-datasets for better organization and analysis.
- Evaluated the effectiveness of sub-datasets in addressing specific aspects of the project.

Incorporation of Libraries for Feature Engineering:

- Reviewed the libraries integrated into both projects for feature extraction and manipulation.

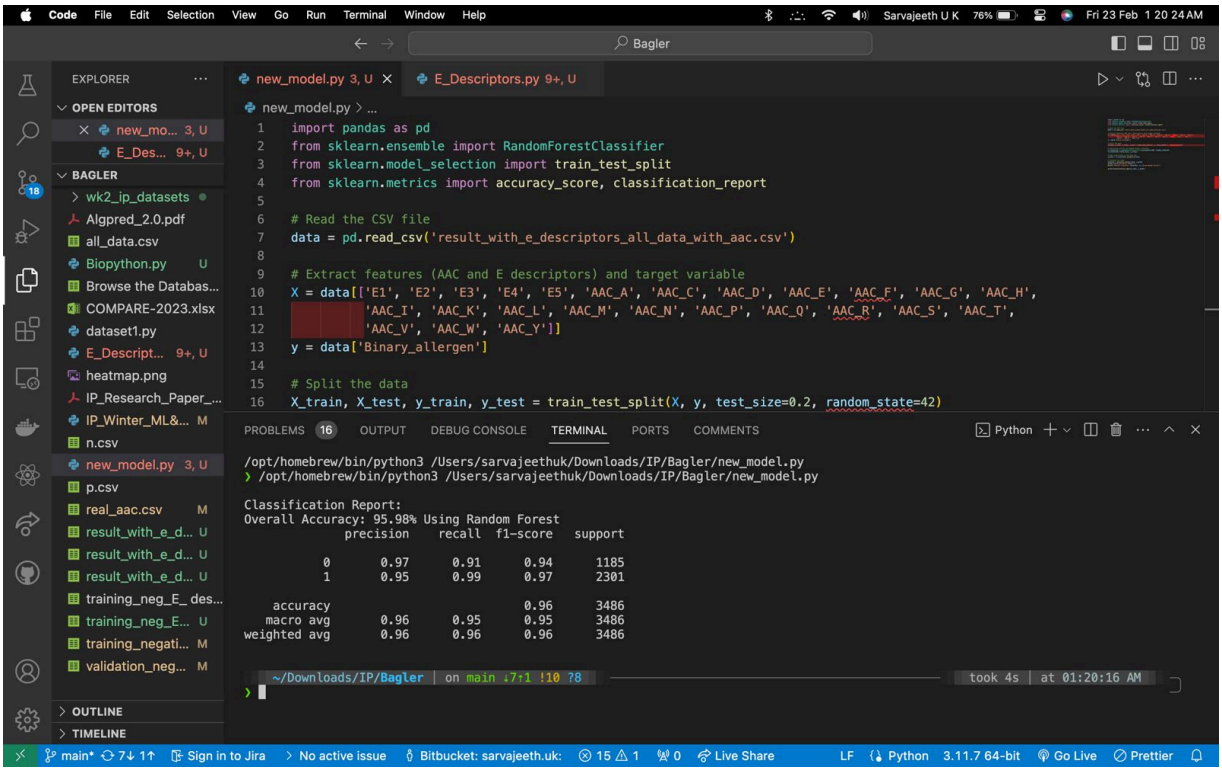
- Assessed the impact of library selection on the quality and efficiency of feature engineering.

During our discussion, we also explored the option of updating data from the website. However, Sir advised against doing so at this time due to the lack of clarity regarding potential improvements or decrements in accuracy. Consequently, we opted not to update the data and are continuing to utilize the previously employed dataset.

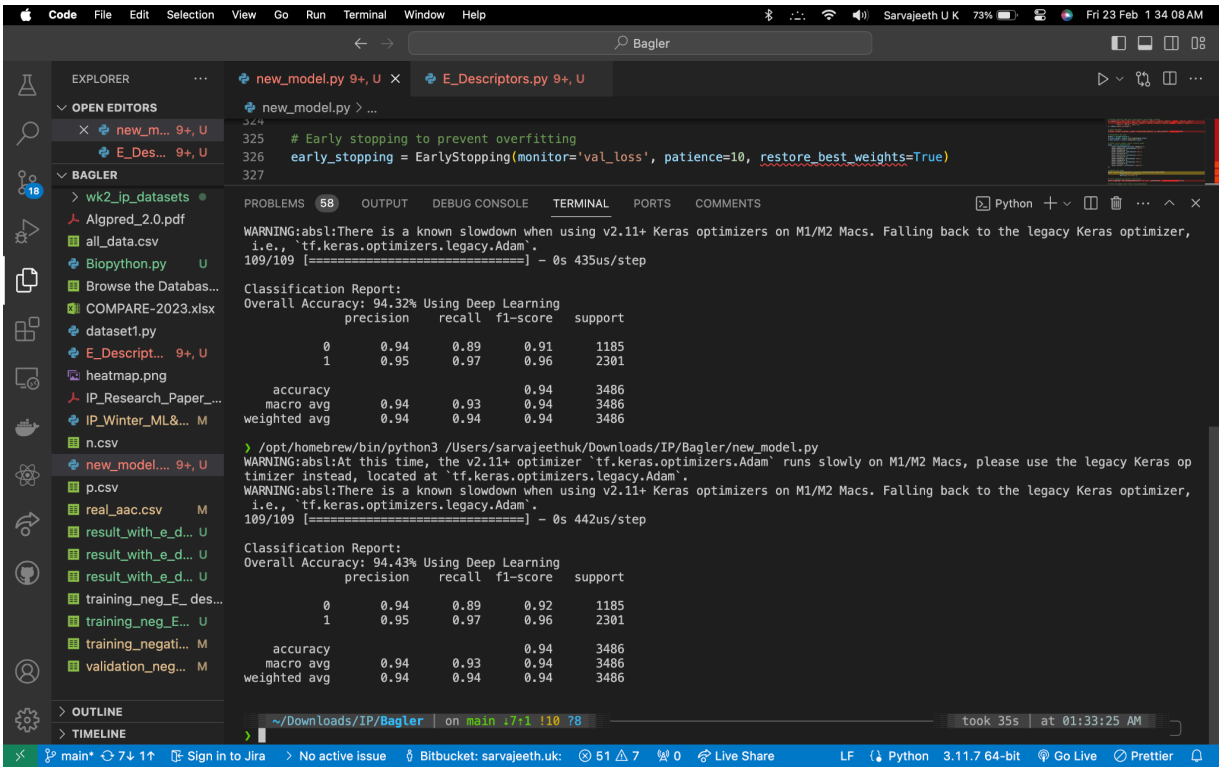
Progress

We found the 5 E descriptors and added it to the ML model with Pfeatures, here is the results we got:

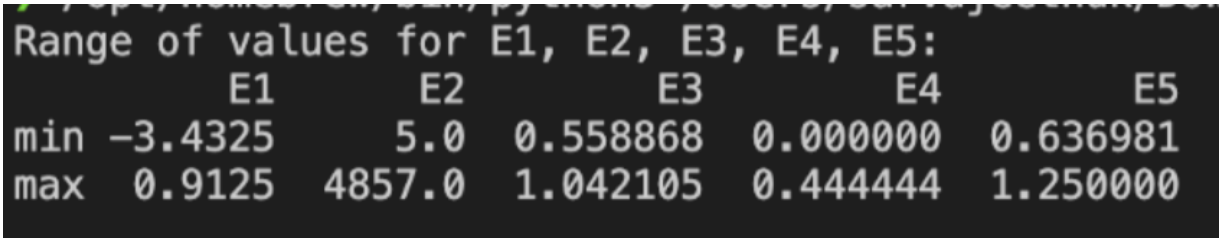
1. **Random Forest - ML Model - 96%**



2. **DL Model using neural network - 94.5%**



Range of E Descriptors we Calculated:



We found the rest of the E Descriptors:

E1: Hydrophilic Nature of Peptides

- Calculation: Average hydrophobicity of amino acids in the peptide sequence.

- Function: Indicates the overall hydrophilicity or hydrophobicity of the peptide. Hydrophobic peptides tend to interact more with lipid bilayers, while hydrophilic peptides are more soluble in aqueous solutions.

E2: Length of Peptides

- Calculation: Simply the length of the peptide sequence.
- Function: Provides information about the size of the peptide, which can influence its stability, interactions, and biological activity.

E3: Tendency for Helical Formation-
https://www.cell.com/fulltext/S0006-3495%2898%2977529-0

- Calculation: Average helical propensity of amino acids in the peptide sequence.
- Function: Indicates the likelihood of the peptide to adopt a helical secondary structure. Helical peptides are common in protein-protein interactions and membrane-spanning regions.

E4: Abundance and Distribution of Positively Charged Amino Acids

- Calculation: Ratio of positively charged amino acids (Arginine and Lysine) to the total number of amino acids.
- Function: Reflects the net positive charge of the peptide, which can influence its interaction with negatively charged molecules such as nucleic acids or cell membranes.

E5: Tendency for β Strand Formation

- Calculation: Average strand propensity of amino acids in the peptide sequence.
- Function: Indicates the propensity of the peptide to form β-strands, which are common secondary structures in proteins and play roles in stability and interaction interfaces.

E6: Disulfide Bond Formation

- Calculation: Count of cysteine residues in the peptide sequence, divided by 2 (assuming each disulfide bond involves two cysteine residues).
- Function: Reflects the potential for disulfide bond formation, which can stabilize the peptide structure and affect its biological activity.

E7: Stability from Pepsin (pH 1.0)

- Calculation: Example stability metric (molecular weight in this case).

- Function: Represents the stability of the peptide under extreme conditions (pH 1.0 in this case). Actual stability metrics may vary depending on experimental conditions.

E8: Thermal Stability

- Calculation: Thermal stability calculated using external libraries or methods (e.g., Pfeature).
- Function: Indicates the resistance of the peptide to thermal denaturation, which is important for applications requiring stability at high temperatures.

E9: Susceptibility at Low pH - Link

- Calculation: Ratio of low pH-susceptible amino acids (Aspartic Acid and Glutamic Acid) to the total number of amino acids.
- Function: Reflects the susceptibility of the peptide to acidic conditions, which can influence its stability and function in acidic environments.

E10: Cross-Reactivity - Link

- Calculation: Average cross-reactivity score of amino acids in the peptide sequence.
- Function: Indicates the likelihood of the peptide to cross-react with other molecules, such as antibodies or receptors, based on the propensity of its constituent amino acids.

E11: Presence of Motifs

- Calculation: Count of a specific motif ('xyz') in the peptide sequence, normalized by the length of the sequence.
- Function: Indicates the presence of specific sequence motifs within the peptide, which may have functional or structural significance.

Doubts:

1. E7 - PFeature Lib (Thermal stability)
2. E11- Motifs Name
3. Check Cross Reactivity
4. How much weight should we assign to particular feature - 5x
5. E12

Summary

BioPython vs Pfeature comparison

1. Why AAC - Auto Cross covariance
2. Why you got particular ranges of E-descpriptor
3. Stability from Pepsin
4. Electro +ve & -ve

E12 - use result from model as a feature - 5x
Presence of Allergenicity

```
# Define hydrophobicity scale (Kyte-Doolittle scale as an example)
   hydrophobicity_scale = {'m': 1.9, 'l': 1.3, 'i': 1.3, 'f': 2.8, 'v': 1.0, 's':
-0.8, 'p': -1.6, 't': -0.7, 'a': 0.5, 'y': 2.8, 'h': -3.2, 'q': -3.5, 'n': -3.5, 'k':
-3.9, 'd': -3.5, 'e': -3.5, 'c': 2.5, 'w': -0.9, 'r': -4.5, 'g': 0.4}

   # Define helical propensity scale
   helical_propensity_scale = {'m': 1.1, 'l': 0.9, 'i': 1.1, 'f': 1.2, 'v': 1.1, 's':
0.8, 'p': 1.1, 't': 0.8, 'a': 0.8, 'y': 0.7, 'h': 0.7, 'q': 1.0, 'n': 1.0, 'k': 0.6,
'd': 0.9, 'e': 1.0, 'c': 0.7, 'w': 1.3, 'r': 0.6, 'g': 0.6}

   # Define strand propensity scale
   strand_propensity_scale = {'m': 1.2, 'l': 1.4, 'i': 1.4, 'f': 1.2, 'v': 1.3, 's':
1.1, 'p': 0.7, 't': 1.2, 'a': 0.7, 'y': 1.0, 'h': 0.9, 'q': 1.0, 'n': 1.2, 'k': 1.2,
'd': 1.3, 'e': 1.3, 'c': 0.8, 'w': 1.0, 'r': 0.7, 'g': 0.5}

cross_reactivity_scores = {'a': 0.2, 'r': 0.8, 'n': 0.5, 'd': 0.7, 'c': 0.3, 'q': 0.5,
'e': 0.7, 'g': 0.2, 'h': 0.6,
                           'i': 0.4, 'l': 0.4, 'k': 0.7, 'm': 0.5, 'f': 0.6, 'p':
0.4, 's': 0.3, 't': 0.3, 'w': 0.7,
                           'y': 0.6, 'v': 0.3}
```

1. **Hydrophobicity Scale (Kyte-Doolittle scale)**:
    a. Reference: Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. Journal of molecular biology, 157(1), 105-132.
    b.
    c. Description: The Kyte-Doolittle scale assigns hydrophobicity scores to amino acids, where higher scores indicate higher hydrophobicity.
    d.
2. **Helical Propensity Scale**:
    a. Reference: Chou, P. Y., & Fasman, G. D. (1978). Empirical predictions of protein conformation. Annual review of biochemistry, 47(1), 251-276.
    b.
    c. Description: The helical propensity scale assigns scores to amino acids based on their propensity to form alpha helices in protein secondary structures.
    d.
3. **Strand Propensity Scale**:
    a. Reference: Stroud, R. M., & Fauman, E. B. (1983). Dynamic processes in protein crystal structures. Accounts of Chemical Research, 16(12), 398-404.
    b.
    c. Description: The strand propensity scale assigns scores to amino acids based on their propensity to form beta strands in protein secondary structures.

Links-
https://www.cell.com/fulltext/S0006-3495%2898%2977529-0 (E3)

https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Hydrophobicity_scales.html (E1) Hydrophobicity scales

https://www.pnas.org/doi/full/10.1073/pnas.96.16.9074 (E5 Beta strands)

1. **Report when we took E6-E11:**

```
Classification Report:
Overall Accuracy: 94.49% Using Deep Learning
              precision    recall  f1-score   support

           0       0.92      0.91      0.92      1185
           1       0.96      0.96      0.96      2301

    accuracy                           0.94      3486
   macro avg       0.94      0.94      0.94      3486
weighted avg       0.94      0.94      0.94      3486
```

2. Report when we took E1-E5:

```
Classification Report:
Overall Accuracy: 94.41% Using Deep Learning
              precision    recall  f1-score   support

           0       0.92      0.91      0.92      1185
           1       0.95      0.96      0.96      2301

    accuracy                           0.94      3486
   macro avg       0.94      0.94      0.94      3486
weighted avg       0.94      0.94      0.94      3486
```

3. Report when we took E1-E11:

```
Classification Report:
Overall Accuracy: 94.09% Using Deep Learning
              precision    recall  f1-score   support

           0       0.92      0.91      0.91      1185
           1       0.95      0.96      0.96      2301

    accuracy                           0.94      3486
   macro avg       0.94      0.93      0.93      3486
weighted avg       0.94      0.94      0.94      3486
```

When we took E12 also, the accuracy reduced to 33.99%:

```
Classification Report (with extended features):
Overall Accuracy: 33.99% Using Deep Learning
/opt/homebrew/lib/python3.11/site-packages/sklearn/metri
e ill-defined and being set to 0.0 in labels with no pre
  _warn_prf(average, modifier, msg_start, len(result))
/opt/homebrew/lib/python3.11/site-packages/sklearn/metri
e ill-defined and being set to 0.0 in labels with no pre
  _warn_prf(average, modifier, msg_start, len(result))
/opt/homebrew/lib/python3.11/site-packages/sklearn/metri
e ill-defined and being set to 0.0 in labels with no pre
  _warn_prf(average, modifier, msg_start, len(result))
              precision    recall  f1-score   support

           0       0.34      1.00      0.51      1185
           1       0.00      0.00      0.00      2301

    accuracy                           0.34      3486
   macro avg       0.17      0.50      0.25      3486
weighted avg       0.12      0.34      0.17      3486
```

Feature comparison : E descriptor vs P feature
(i) Pfeature: AAC feature's come under Pfeatures

(ii) E descriptor : E1…E12

(iii) Normalization techniques: Xnorm= (X- Xmin)/(Xmax- Xmin) , X max is the max value of the E descriptor , and X min is the min value of the E

Model with Only E descriptors

1. **Accuracy with only E-Descriptors Random Forest - 96.21%**

```
sarvajeethuk@Sarvajeeths-MacBook-Air Bagler % /usr/loca
Accuracy: 96.21%
              precision    recall  f1-score   support

           0       0.97      0.92      0.94      1185
           1       0.96      0.99      0.97      2301

    accuracy                           0.96      3486
   macro avg       0.96      0.95      0.96      3486
weighted avg       0.96      0.96      0.96      3486
```

2. **Using Hyper Parameters - 93.14%**

```
/Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/site-pac
g: Do not pass an `input_shape`/`input_dim` argument to a layer. When usir
 object as the first layer in the model instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
Epochs: 10, Optimizer: adam, Activation: relu, Test Accuracy: 91.59%
Epochs: 10, Optimizer: adam, Activation: tanh, Test Accuracy: 90.76%
Epochs: 10, Optimizer: sgd, Activation: relu, Test Accuracy: 88.07%
Epochs: 10, Optimizer: sgd, Activation: tanh, Test Accuracy: 86.78%
Epochs: 20, Optimizer: adam, Activation: relu, Test Accuracy: 92.37%
Epochs: 20, Optimizer: adam, Activation: tanh, Test Accuracy: 92.23%
Epochs: 20, Optimizer: sgd, Activation: relu, Test Accuracy: 89.67%
Epochs: 20, Optimizer: sgd, Activation: tanh, Test Accuracy: 87.44%
Epochs: 30, Optimizer: adam, Activation: relu, Test Accuracy: 93.14%
Epochs: 30, Optimizer: adam, Activation: tanh, Test Accuracy: 92.86%
Epochs: 30, Optimizer: sgd, Activation: relu, Test Accuracy: 91.25%
Epochs: 30, Optimizer: sgd, Activation: tanh, Test Accuracy: 88.44%
```

3. **Accuracy using ANN - 91.85%**

```
Accuracy: 91.85%
              precision    recall  f1-score   support

           0       0.94      0.82      0.87      1185
           1       0.91      0.97      0.94      2301

    accuracy                           0.92      3486
   macro avg       0.92      0.89      0.91      3486
weighted avg       0.92      0.92      0.92      3486
```

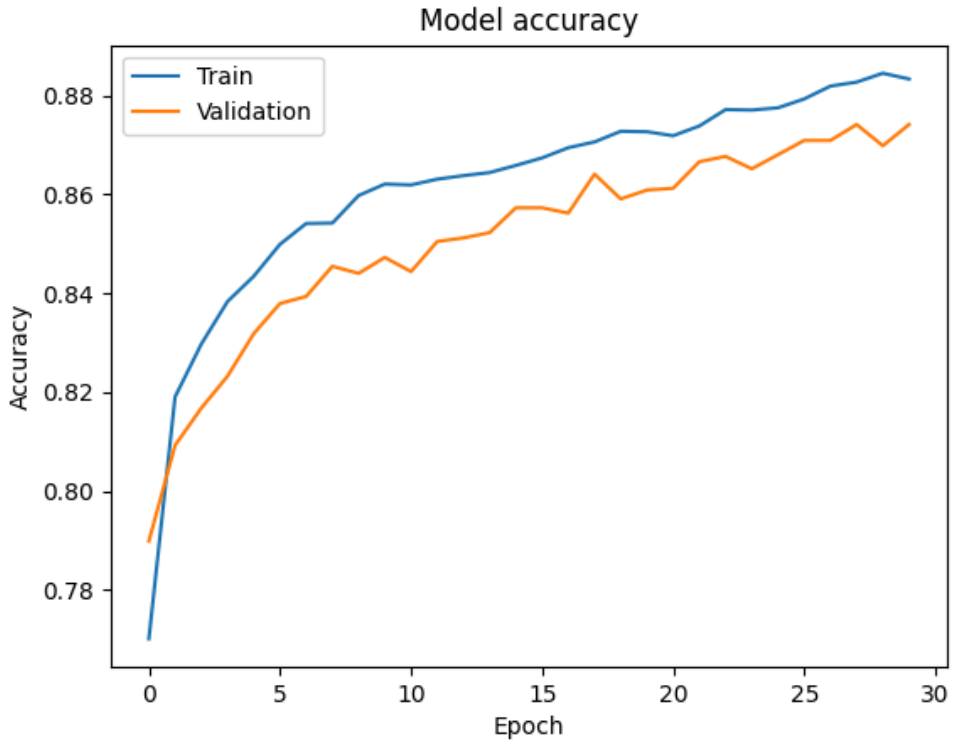
Table with the Accuracy of the Models

Model	Features	Accuracy (%)
Random Forest (ML)	Pfeatures	95.00
Logistic Regression (ML)	Pfeatures	81.00
Sequential NN (DL)	Pfeatures	90.00
Hyper-Parameter Tuning (DL)	Pfeatures	91.00
Random Forest (ML)	Pfeatures + E-Descriptors	96.00
Sequential NN (DL)	Pfeatures + E-Descriptors	94.50
Random Forest (ML)	E-Descriptors	96.21
Hyper-Parameter Tuning (DL)	E-Descriptors	93.14

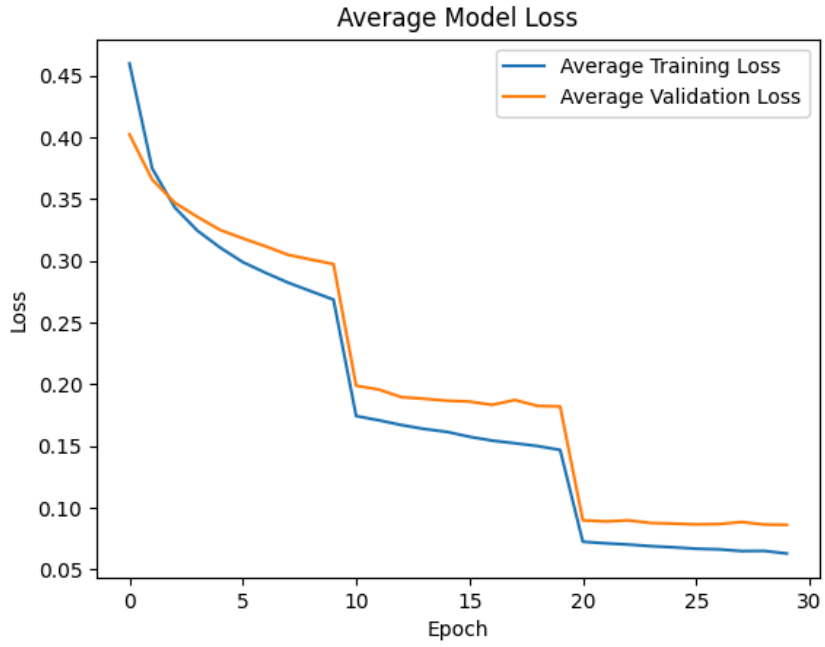
OverFit and UnderFit

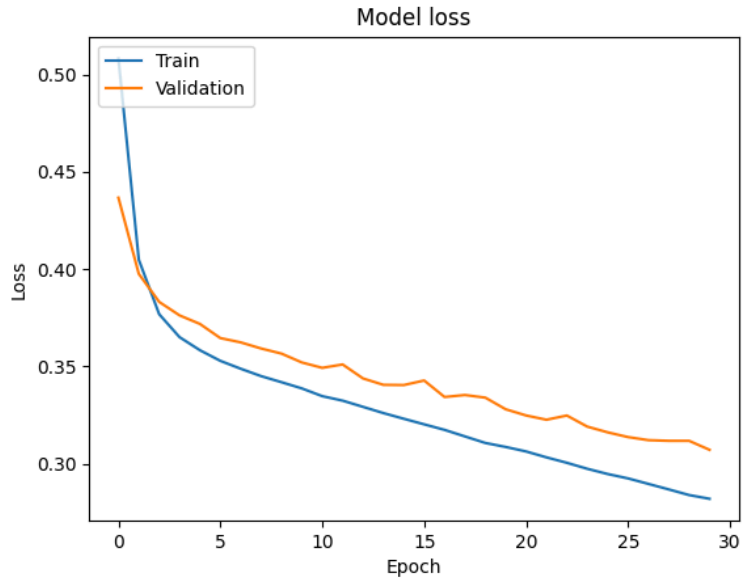
Average Overfit Percentage: 2.01%
Average Underfit Percentage: 0.00%

1. **Plot of Accuracy of Train vs Validation**



2. Average Model Loss





3. Confusion Matrix