

Identification of Disease-Associated Genes using ML

Nisarg Upadhyaya [19CS30031]
Neha Dalmia [19CS30055]



Dataset

- The dataset from DisGeNET comprises gene-disease associations for 21,671 genes, totaling 32,61,324 interactions.
- Gene sequences, including all protein isoforms, are sourced from NCBI.

Statistic	Count
Total sequences	102827
Disease associated sequences	59967
Non disease associated sequences	42860
MeSH disease classes	26



Feature extraction (1/2)


The following features were included using FEPS:

1. Incorporating spatial amino acid relationships through **SOC features** enhances ML algorithm performance in tasks like protein classification, prediction, and function annotation.
2. **Autocorrelation descriptors** (ACDs) capture patterns and correlations in amino acid properties along the sequences



Feature extraction (2/2)

3. **Entropy, relative entropy, and gain**
4. **Composition, transition, and distribution** representing amino acid frequencies, adjacent changes, and spatial patterns along sequences, respectively, also called **AADs (amino acid descriptors)**.
5. **Conjoint triad** method groups amino acids into overlapping triplets, encoding them based on physicochemical properties to capture spatial information and structural motifs.



1094 features extracted for each sequence. Next steps?

Now for each sequence we also have information from DisGeNET about its disease associations.

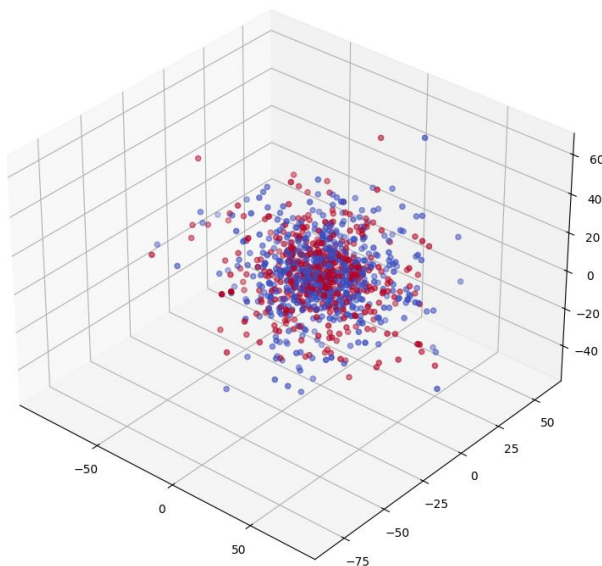
So we try to create models which can learn these associations

1. Traditional ML methods for binary classification
2. Neural network for multi label classification

Before training, a small digression!

Traditional ML methods don't handle high dimensional data well.
Hence, we incorporate PCA to reduce the dimensionality of our data.

1094 -> 43/83
features!



e.g., first three
principal
components for
the amino acid
descriptors



Precision - 0.73

Recall - 0.70

F1 Score - 0.67

Accuracy - 0.70

Traditional ML methods for
binary classification

1. **Support Vector Machine**
2. Gradient Boosting
3. Random Forest
4. Ada Boost
5. Quadratic Discriminant
Analysis



Precision - 0.70

Recall - 0.68

F1 Score - 0.65

Accuracy - 0.68

Traditional ML methods for
binary classification

1. Support Vector Machine
2. **Gradient Boosting**
3. Random Forest
4. Ada Boost
5. Quadratic Discriminant Analysis



Precision - 0.97

Recall - 0.97

F1 Score - 0.97

Accuracy - 0.97

Traditional ML methods for
binary classification

1. Support Vector Machine
2. Gradient Boosting
3. Random Forest
4. Ada Boost
5. Quadratic Discriminant Analysis



Precision - 0.63

Recall - 0.63

F1 Score - 0.61

Accuracy - 0.63

**Traditional ML methods for
binary classification**

1. Support Vector Machine
2. Gradient Boosting
3. Random Forest
4. Ada Boost
5. Quadratic Discriminant Analysis



Precision - 0.63

Recall - 0.62

F1 Score - 0.56

Accuracy - 0.62

**Traditional ML methods for
binary classification**

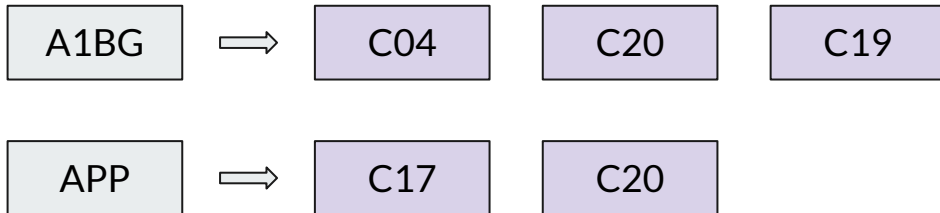
1. Support Vector Machine
2. Gradient Boosting
3. Random Forest
4. Ada Boost
5. Quadratic Discriminant Analysis



Neural Network Architecture (1/3)

Given the high dimensionality of our data, for **multilabel classification** an ideal choice is to adapt a neural network based deep learning approach.

Original way of classification :

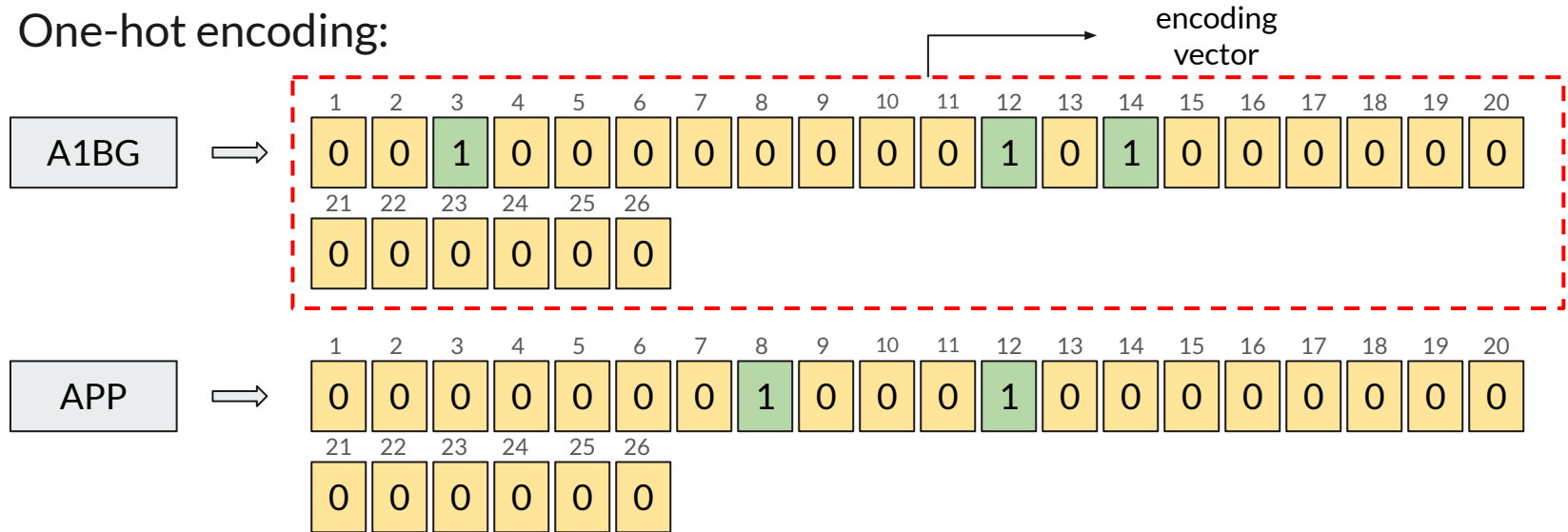


Gene

MeSH Class

Neural Network Architecture (2/3)

One-hot encoding:

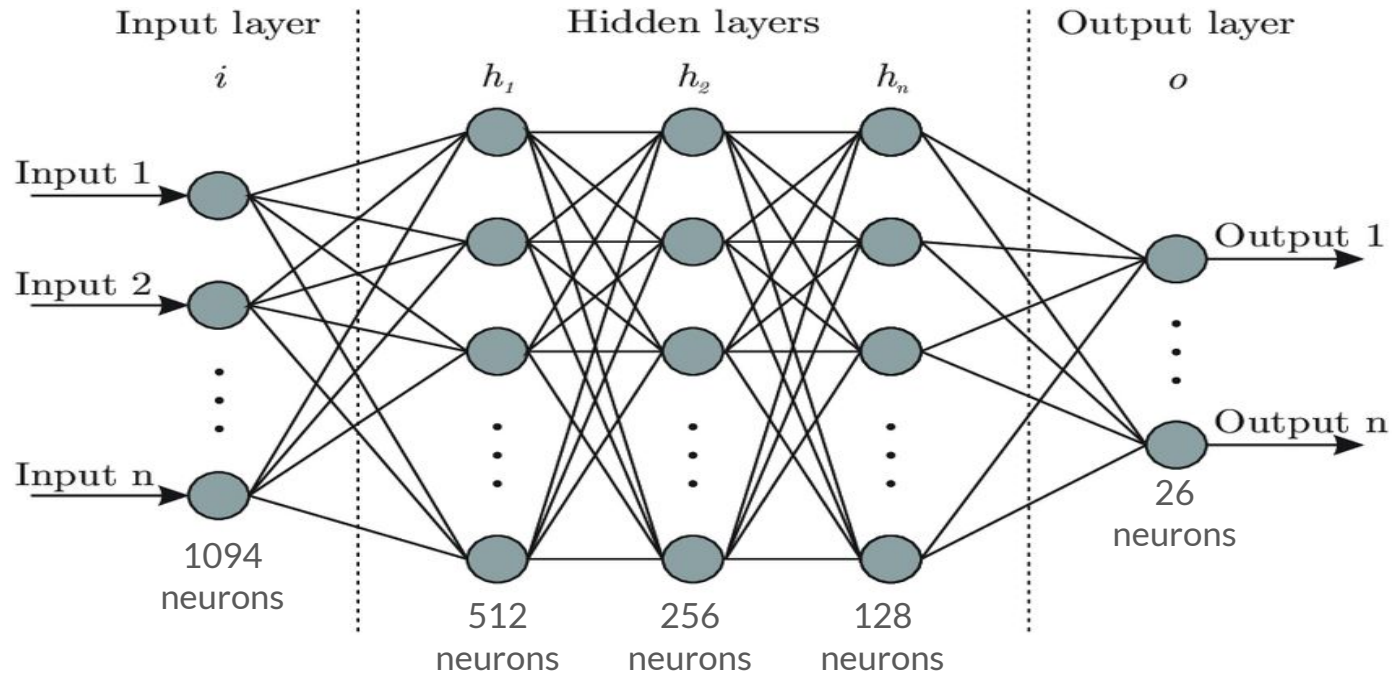


Gene

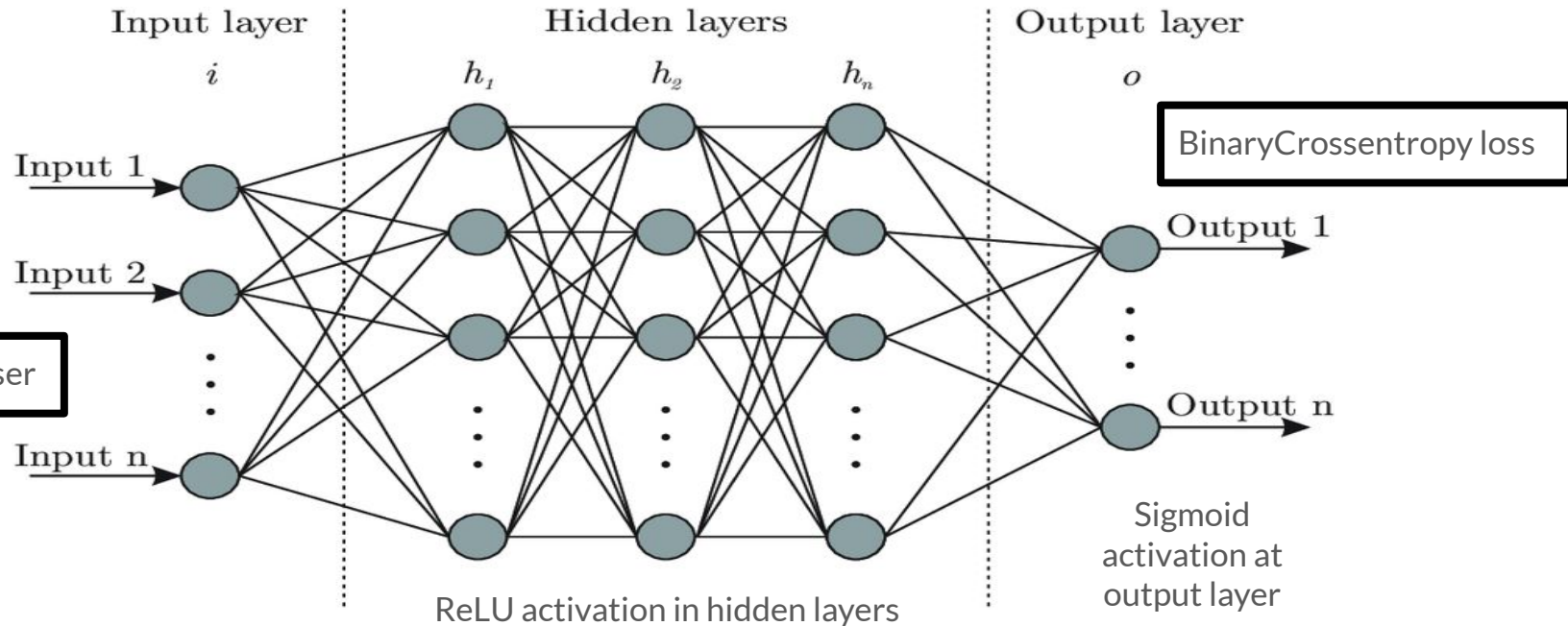
Non presence of Indexed MeSH class corresponding to index

Presence of Indexed MeSH class corresponding to index

Neural Network Architecture (3/3)

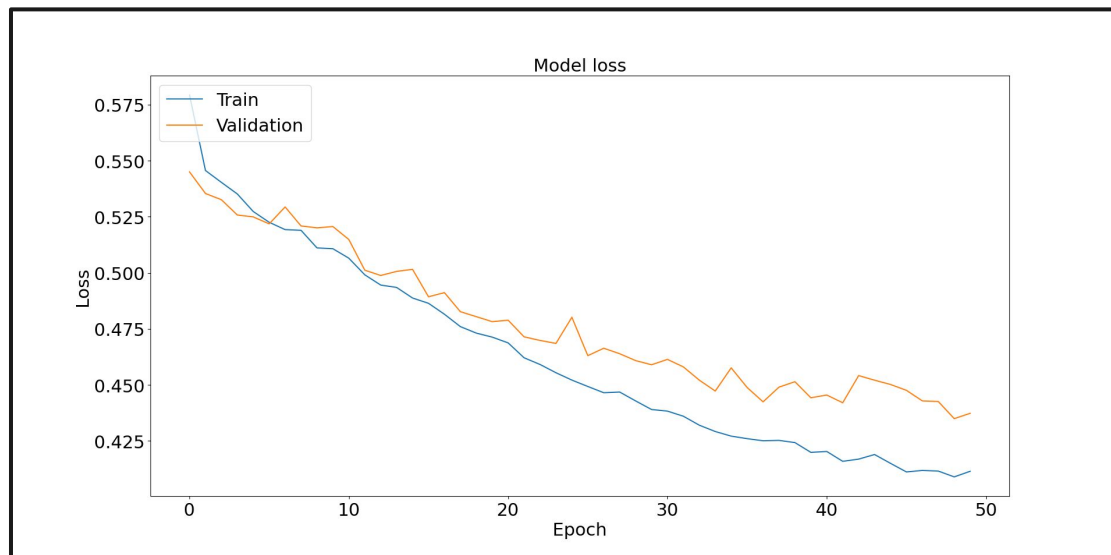


Neural Network Architecture (3/3)



Training

Trained the network for 50 epochs after which the loss converged.



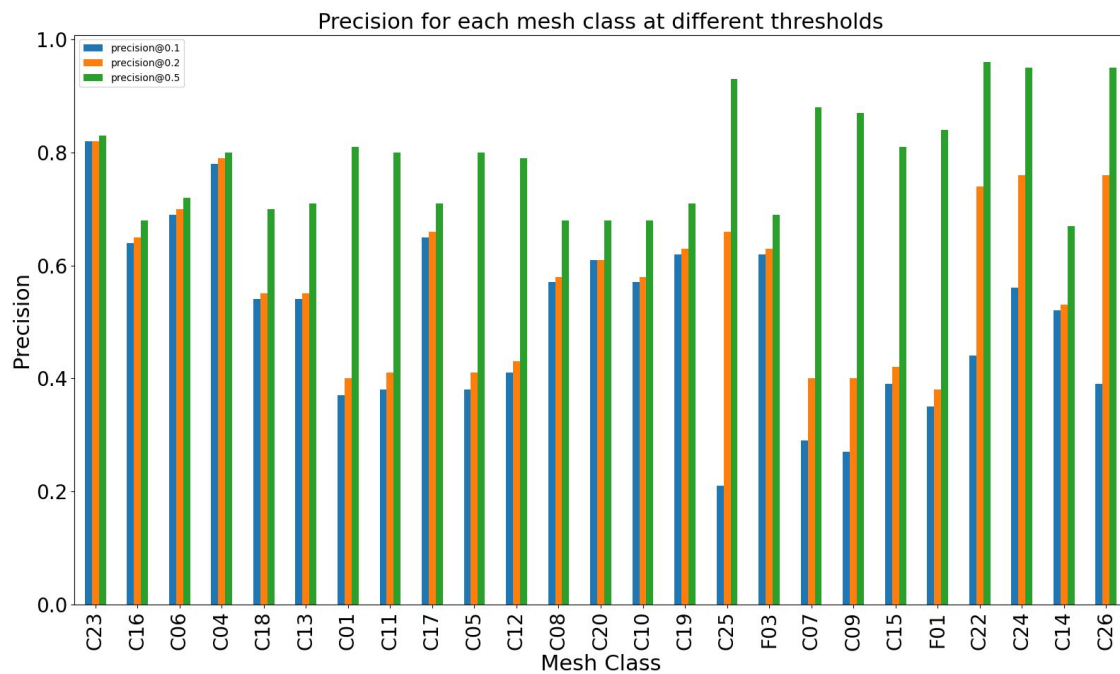


Results

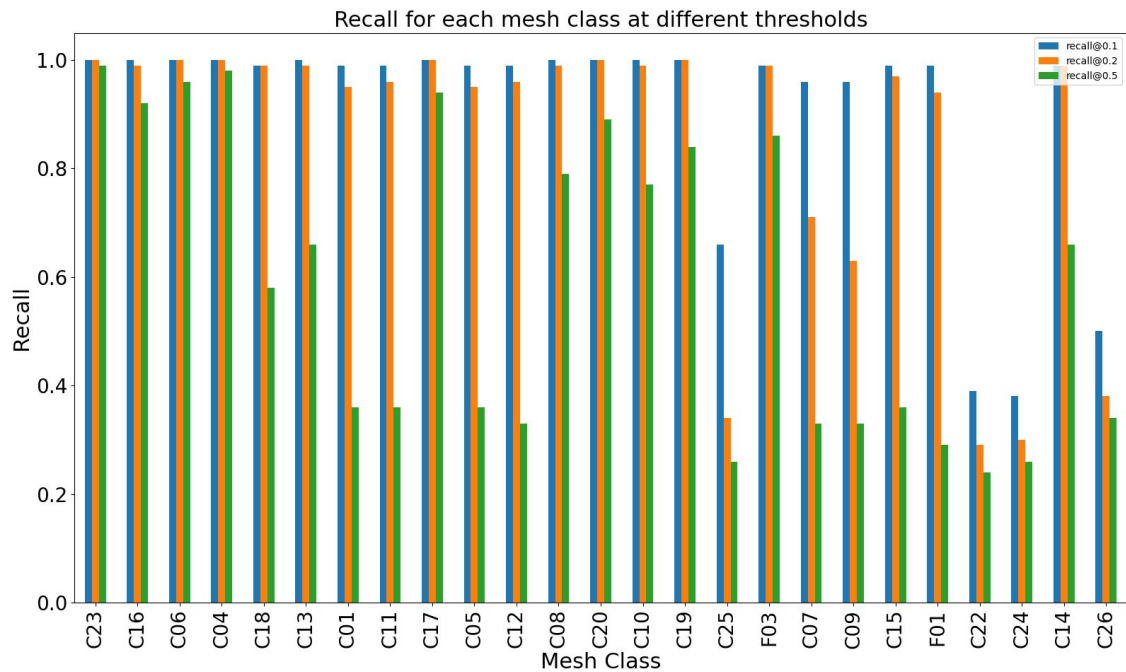
- The trained network outputs probabilities for each of the 26 classes.
- We set different threshold values (0.1, 0.2 and 0.5) to assign a disease class to a gene based on whether the probability for that class is greater than the threshold or not.

Metric (weighted average)	p > 0.1	p > 0.2	p > 0.5
Precision	0.57	0.60	0.75
Recall	0.98	0.96	0.72
F1-Score	0.71	0.72	0.70

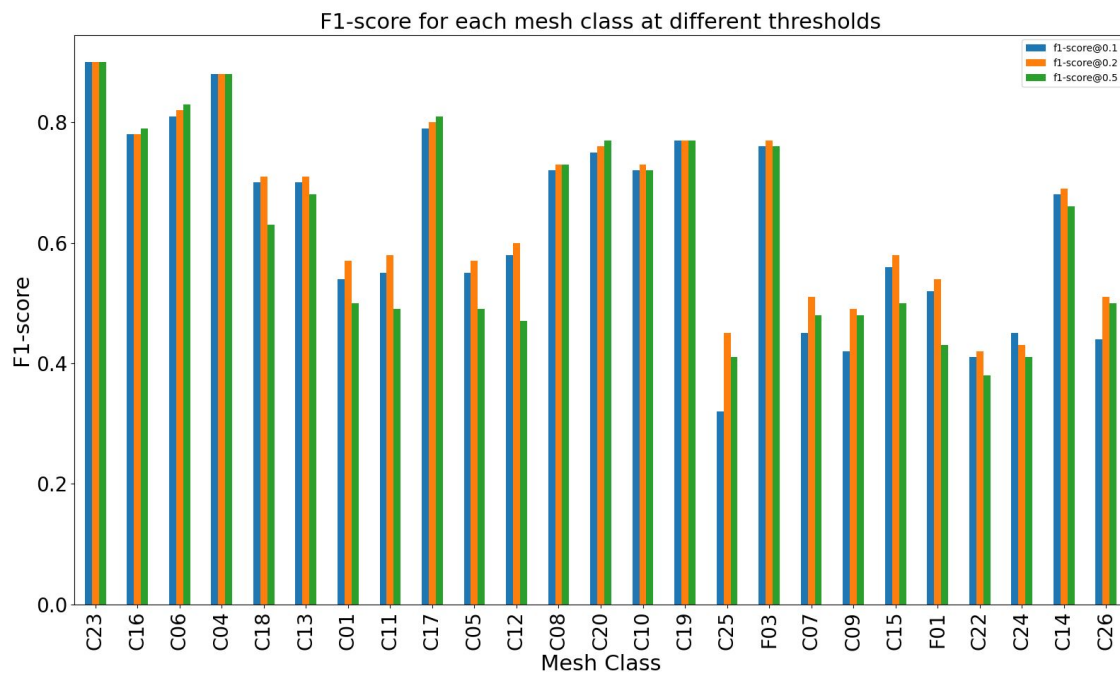
Results



Results



Results



Thank you!