Protein families classification using Machine/Deep Learning

Abishek Adhikari

Motivation/Objective:

- Proteins are the complex biomolecules that plays an important role in living organism.
- Proteins are made up of amino acids; a long chain of chemical units represented by alphabets
- There are 20 most common amino acids that are arranged randomly in a sequence which generates millions of unique proteins
- Traditional experimental methods uses chemical and physical properties to classify protein families, which is complex and time consuming
- This project aims to classify protein families using several Machine/Deep Learning approaches.

Data wrangling

- Protein sequence data is downloaded from Kaggle. (https://www.kaggle.com/shahir/protein-data-set)
- There are two data files, one contains protein sequence whereas other is physical and chemical properties of protein.
- More than 85000 Protein families are present in the data sets with approximately 3900 classes.
- Each features are carefully examine for missing values and the missing instances are removed from the data set.
- After cleaning, approximately 75000 instances are available for the further steps.

Understanding data

Dataset 1

#	Column	Non-Null Count	Dtype
0	structureId	75010 non-null	object
1	classification	75010 non-null	object
2	experimentalTechnique	75010 non-null	object
3	residueCount	75010 non-null	int64
4	resolution	75010 non-null	float64
5	structureMolecularWeight	75010 non-null	float64
6	densityMatthews	75010 non-null	float64
7	densityPercentSol	75010 non-null	float64
8	phValue	75010 non-null	float64
9	sequence	75010 non-null	object
10	label	75010 non-null	int64
11	lengths	75010 non-null	int64

Dataset 2

#	Column	Non-Null Count	Dtype
0	structureId	104813 non-null	object
1	chainId	104812 non-null	object
2	sequence	104812 non-null	object
3	residueCount	104813 non-null	int64
4	macromoleculeType	101336 non-null	object

Amino acids are represented by alphabets

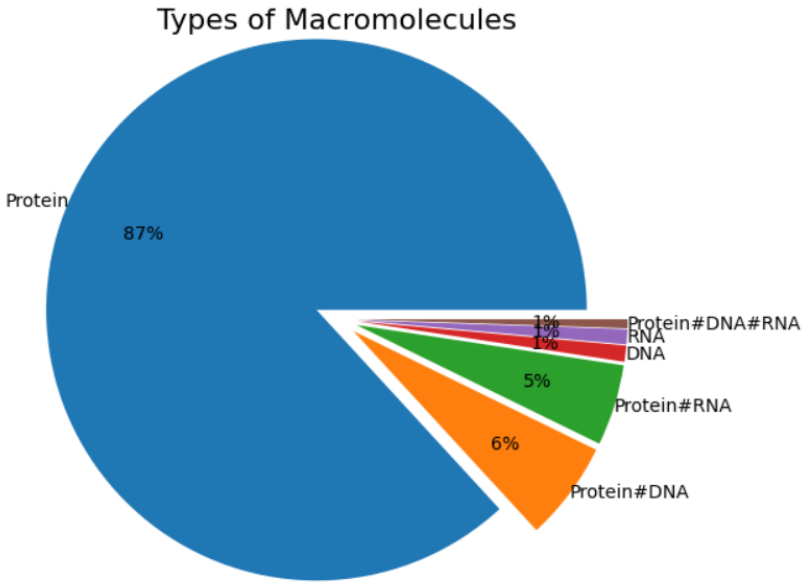
Amino Acid	3-Letter Code	1-Letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid or aspartate	Asp	D
Glutamic acid or glutamate	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

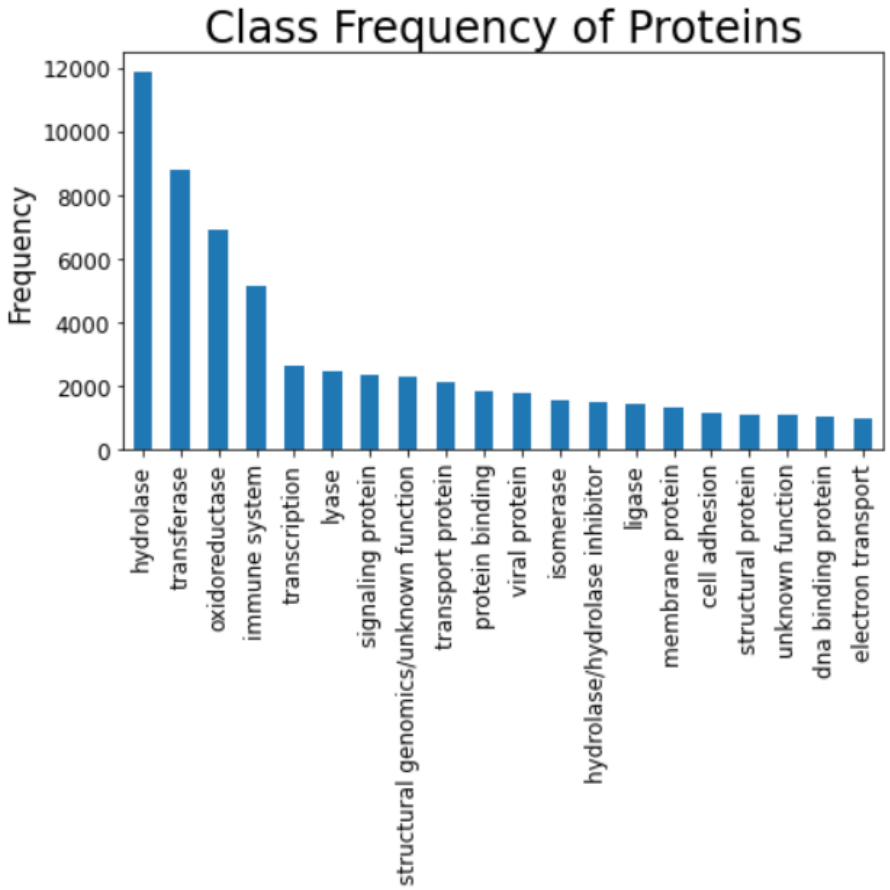
“Hydrolase” family with 286 amino acid units

```
'TYTTRQIGAKNTLEYKV
YIEKDGKPVSAFHDIPLY
ADKENNIFNMVVEIPRWT
NAKLEITKEETLNPIIQD
TKKGKLRFVRNCFPHHGY
IHNYGAFPQTWEDPNVSH
PETKAVGDNEPIDVLEIG
ETIAYTGQVKQVKALGIM
ALLDEGETDWKVIAIDIN
DPLAPKLNDIEDVEKYFP
GLLRATNEWFRIYKIPDG
KPENQFAFSGEAKNKKYA
LDIIKETHDSWKQLIAGK
SSDSKGIDLTNVTLPDTP
TYSKAASDAIPPASLKAD
APIDKSIDKWFFISGSV'
```

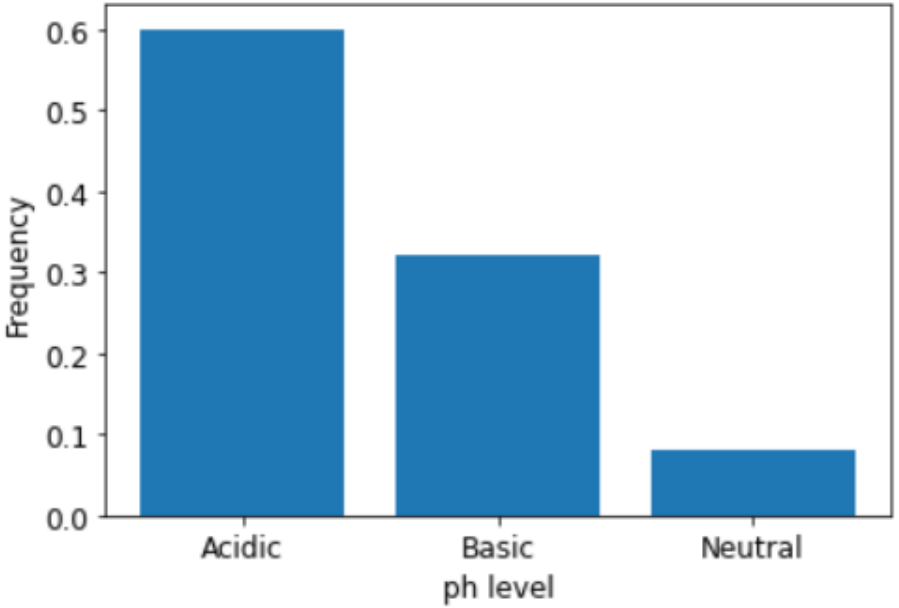
Protein families	Numerical Labels:
'hydrolase'	: 1
'hydrolase/hydrolase inhibitor'	: 13
'immune system'	: 4
'isomerase'	: 12
'ligase'	: 14
'lyase'	: 6

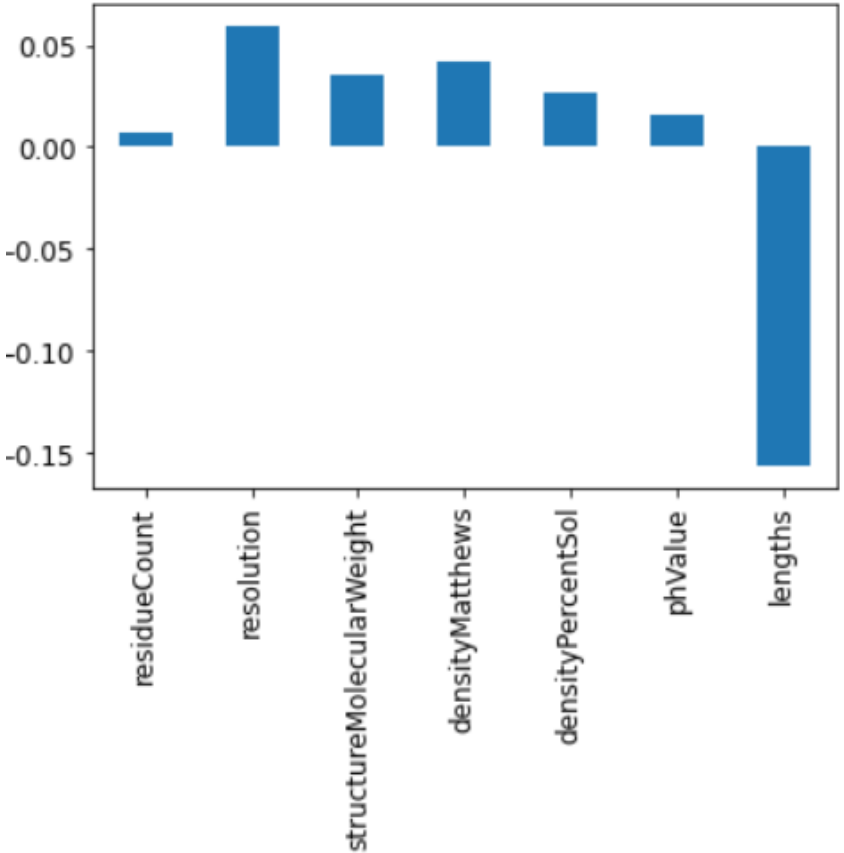
Exploratory Data Analysis





Exploratory Data Analysis





Machine Learning Models

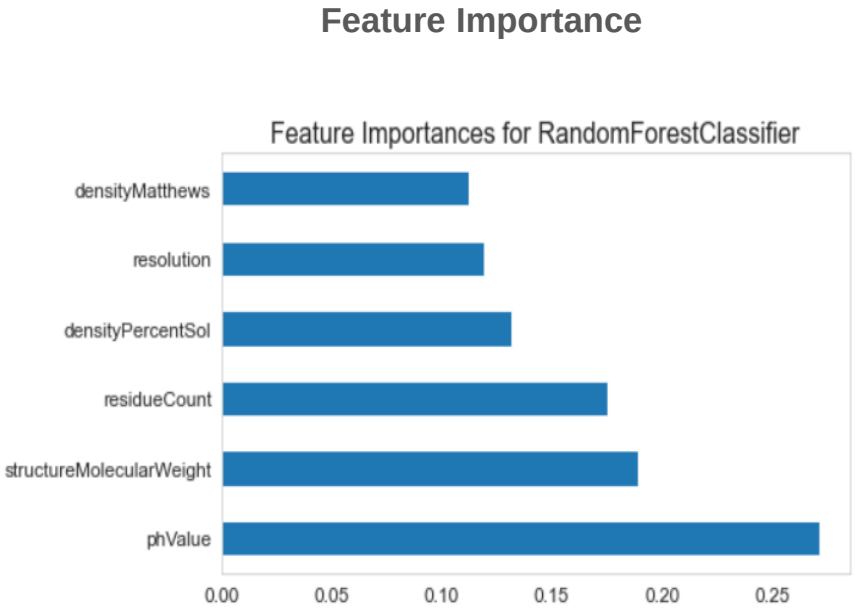
Three ML algorithms (Random Forest, Decision Tree, and KNN) are trained and tested.

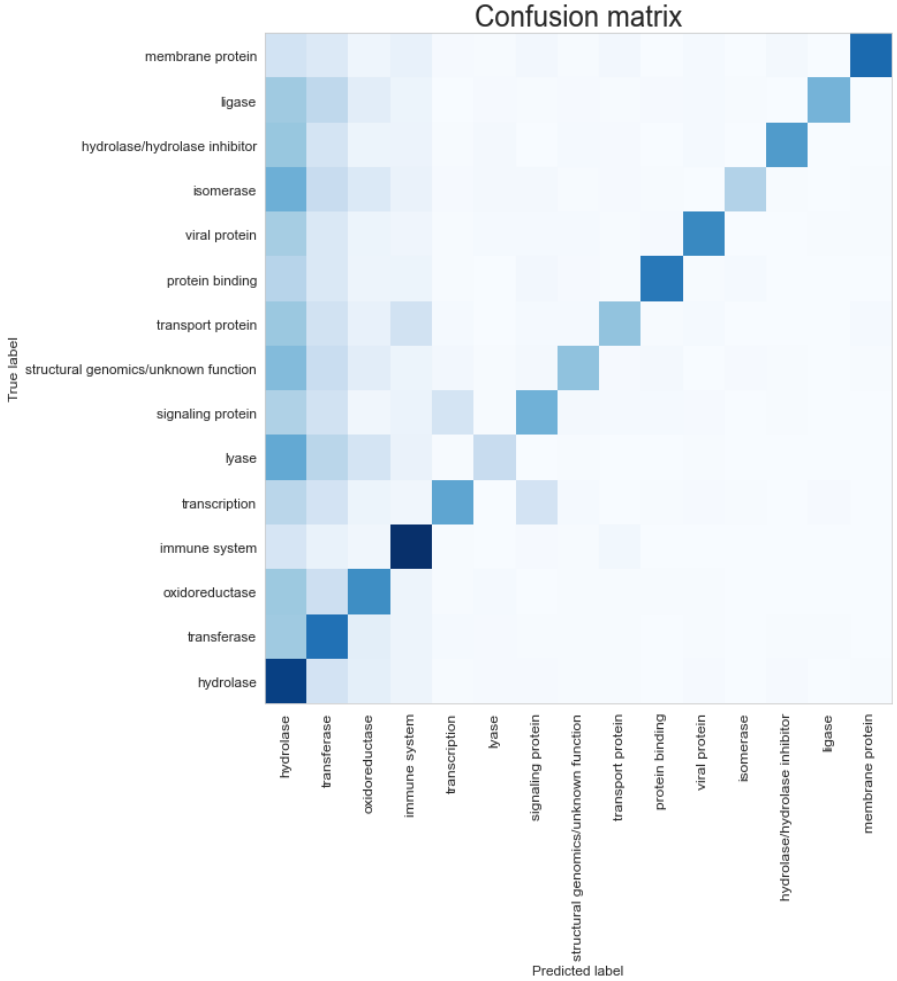
75% data are used to train the model and the remaining 25% are tested

Hyperparameters are tuned in each model

Models	Accuracy	Recall	precision
Decision Tree	0.43	0.43	0.49
Random Forest	**0.52**	**0.52**	**0.57**
KNN	0.28	0.28	0.27

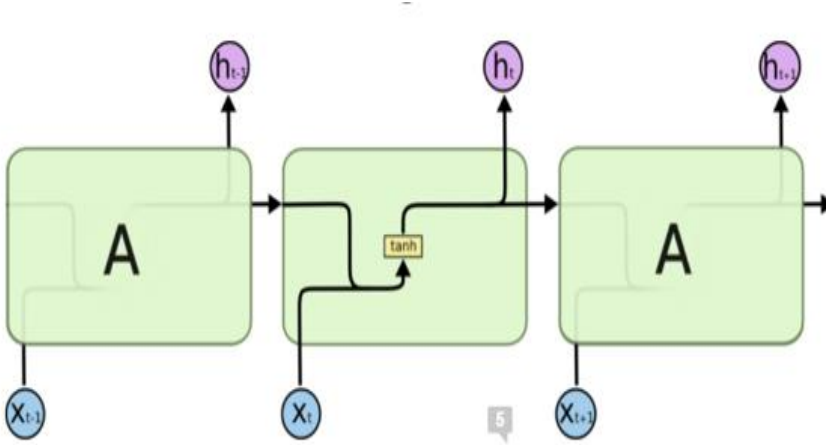
Random Forest





Brief introduction of Sequential Model

Recurring Neural Network (RNN)



The repeating module in a standard RNN contains a single layer.

In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other.

An RNN can be thought of as multiple copies of the same network, each passing a message to a successor.
RNNs can learn to use the past information.

Unfortunately, as that gap grows, RNNs become unable to learn to connect the information.

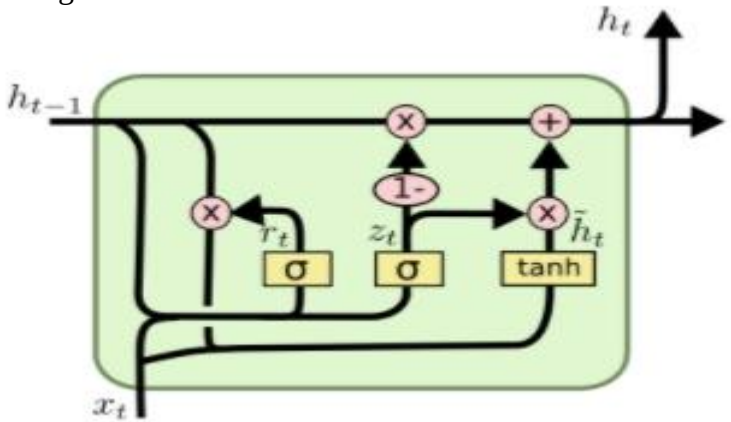
Long Short-Term Memory (LSTM)

"LSTMs" – are a special kind of RNN, capable of learning long-term dependencies

Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

The key to LSTMs is the cell state, the horizontal line running through the top of the diagram.

The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates



Deep Learning Models (Model architecture):

LSTM

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 1, 256)	5383168
dropout (Dropout)	(None, 1, 256)	0
lstm_1 (LSTM)	(None, 1, 256)	525312
dropout_1 (Dropout)	(None, 1, 256)	0
lstm_2 (LSTM)	(None, 256)	525312
dense (Dense)	(None, 15)	3855

Total params: 6,437,647
Trainable params: 6,437,647
Non-trainable params: 0

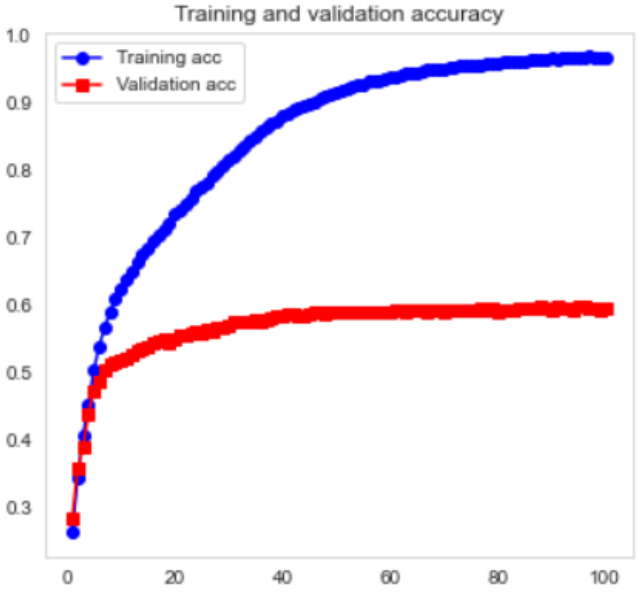
BiDirectional LSTM

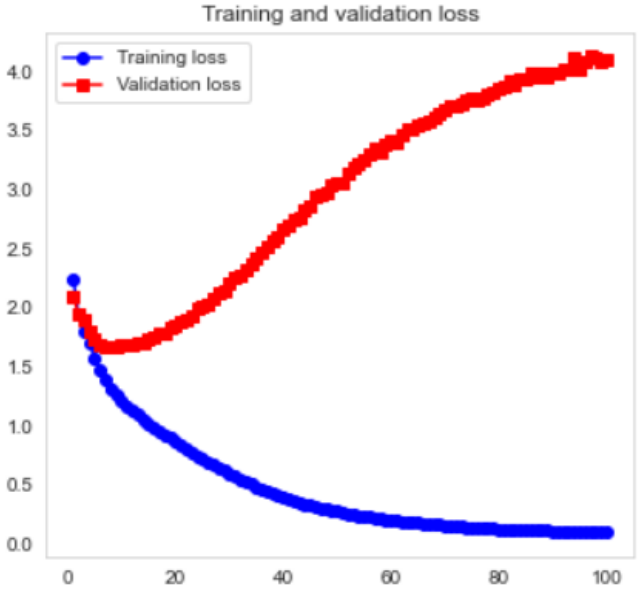
Layer (type)	Output Shape	Param #
bidirectional (Bidirectional	(None, 1, 512)	10766336
dropout_2 (Dropout)	(None, 1, 512)	0
bidirectional_1 (Bidirection	(None, 1, 512)	1574912
dropout_3 (Dropout)	(None, 1, 512)	0
bidirectional_2 (Bidirection	(None, 512)	1574912
dense_1 (Dense)	(None, 15)	7695

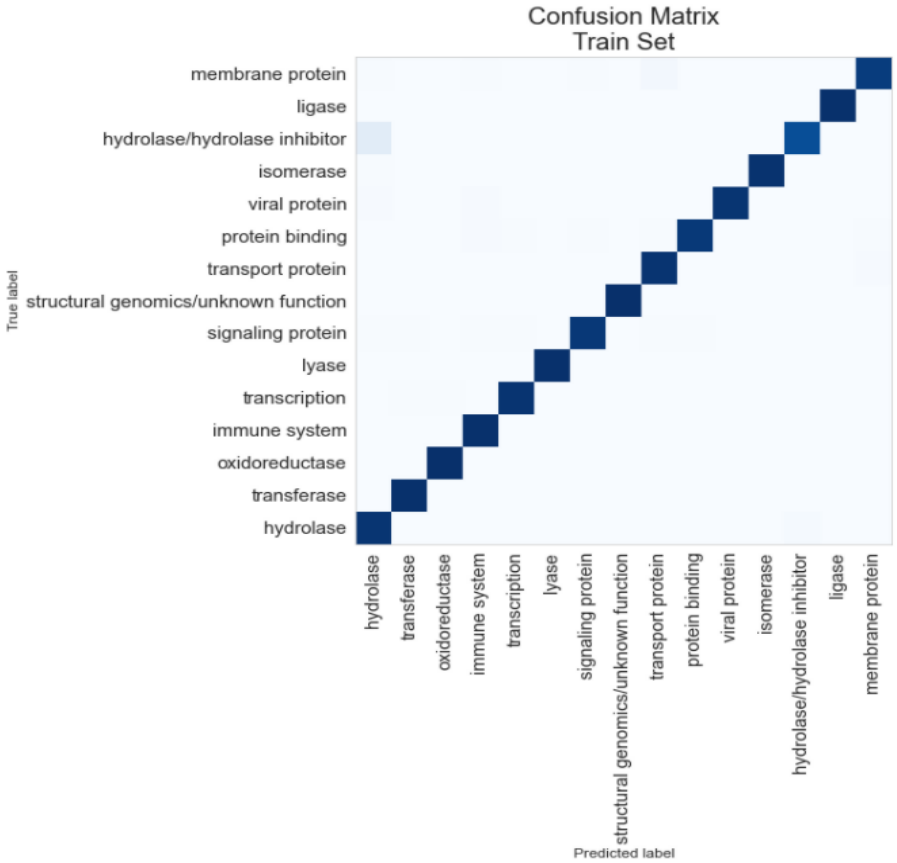
Total params: 13,923,855
Trainable params: 13,923,855
Non-trainable params: 0

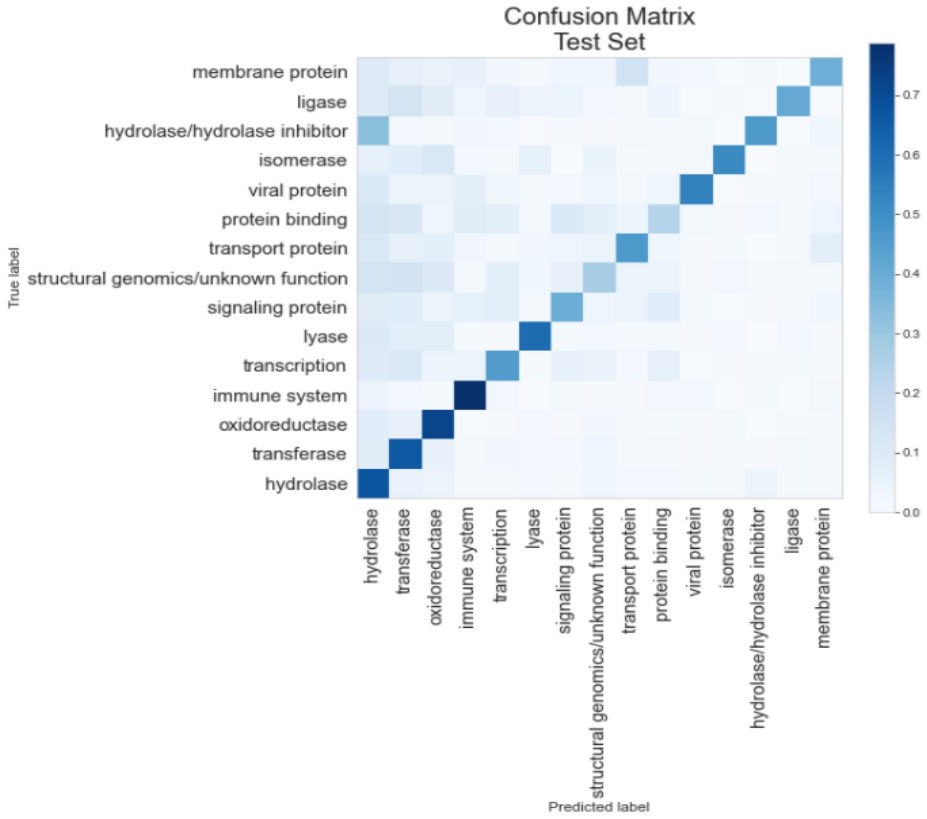
Model evaluation

Models	Accuracy	Recall	precision	AUC
LSTM	0.59	0.59	0.61	0.83
BiLSTM	0.59	0.59	0.60	0.84






Confusion Matrix
Train Set
membrane protein
ligase
hydrolase/hydrolase inhibitor
isomerase
viral protein
protein binding
transport protein
structural genomics/unknown function
signaling protein
lyase
transcription
immune system
oxidoreductase
transferase
hydrolase
True label
hydrolase
transferase
oxidoreductase
immune system
transcription
lyase
signaling protein
structural genomics/unknown function
transport protein
protein binding
viral protein
isomerase
hydrolase/hydrolase inhibitor
ligase
membrane protein
Predicted label


Confusion Matrix
Test Set
membrane protein
ligase
hydrolase/hydrolase inhibitor
isomerase
viral protein
protein binding
transport protein
structural genomics/unknown function
signaling protein
lyase
transcription
immune system
oxidoreductase
transferase
hydrolase
True label
hydrolase
transferase
oxidoreductase
immune system
transcription
lyase
signaling protein
structural genomics/unknown function
transport protein
protein binding
viral protein
isomerase
hydrolase/hydrolase inhibitor
ligase
membrane protein
Predicted label
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Summary and recommendation:

- This projects aims to classify 15 most common protein families using two separate approaches (Machine Learning and Deep Learning).
- Machine learning algorithm uses physical and chemical properties of protein whereas deep learning uses amino acid sequence to predict the protein families.
- Machine learning algorithms are not much useful to classify protein families because of the limited features.
- Both LSTM and BiLSTM are some how useful to classify protein families.

Recommendation:

- More features are need to ML model to improve its predictability.
- More data is needed for LSTM to improve its predictability.

Thank you.