

PROTEIN SUBCELLULAR LOCALIZATION PREDICTION



PROBLEM STATEMENT

Organelle within a cell where a protein is located, providing insights into its function and role in cellular processes. It is a Multi-Label Multi class classification problem with 5 classes named envelope, lumen, plastoglobule, stroma, thylakoid_membrane.

Two evolutionary-based profiles are generated:

- Hidden Markov model (HMM)
- Position-specific scoring matrix (PSSM)

The shapes of HMM and PSSM evolutionary profiles is $(L, 20)$ representing the substitution probabilities. Here, L is the length of the protein sequence.

MODELS

- Using Standardized SXG matrix for Feature Extraction and passing them to Deep neural network.
- Passing Word2Vec Embeddings as input to SXG matrix for feature extraction and passing them through Deep Neural Network.

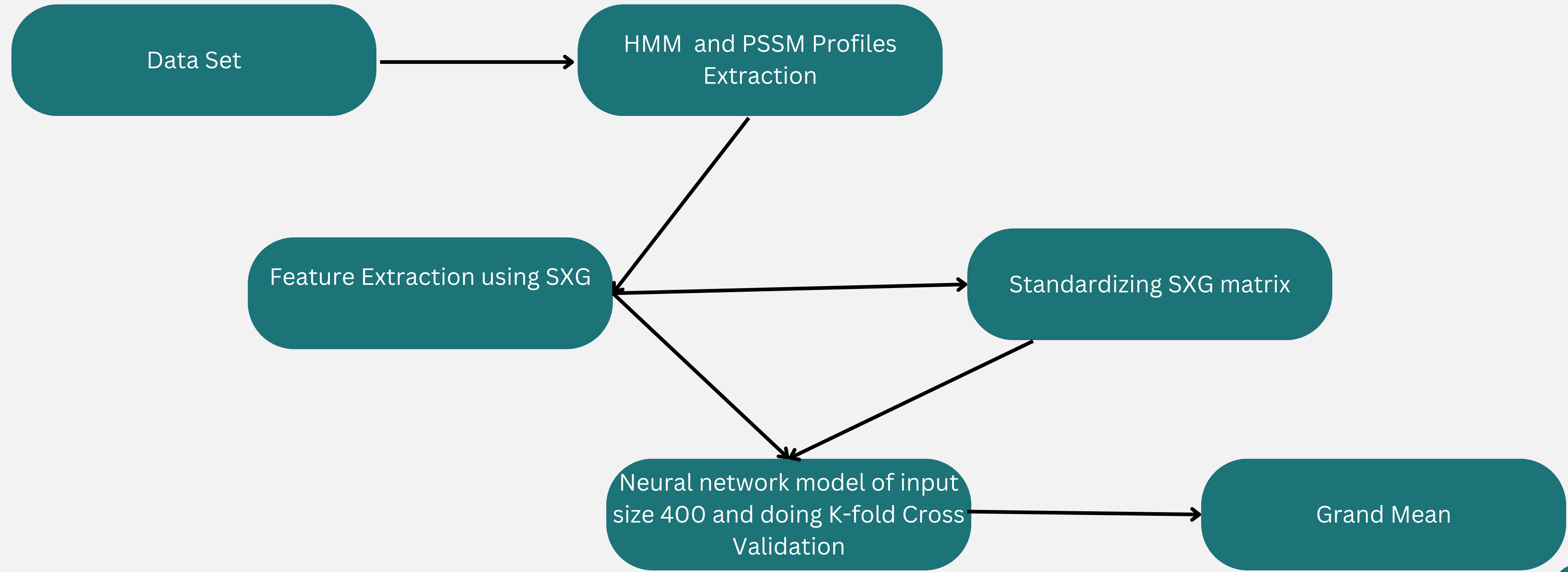




MODEL 1

- Using Standardized SXG matrix for Feature Extraction and passing them to neural network.

FLOW CHART



EVALUATION METRICS

- Overall Validation Accuracy
- Accuracy
- Jaccard Accuracy
- Precision
- Recall
- F1-score

GRAND MEAN = (ACC + VAL ACC + JAC ACC + PRECISION + RECALL + F1)/6

EVALUATION METRICS

Accuracy

Number of correct predictions (where true label matches the predicted label.)

$y\text{-true} = [1\ 0\ 0\ 1\ 0]$ and $y\text{-pred} = [1\ 0\ 1\ 0\ 0]$

Accuracy = 60%

Jaccard Accuracy

Ratio of the intersection (true positives) to the union (true positives + false positives + false negatives) of the true and predicted labels.

$y\text{-true} = [1\ 0\ 0\ 1\ 0]$ and $y\text{-pred} = [1\ 0\ 1\ 0\ 0]$

Jaccard Accuracy = 33.33%

FEATURE EXTRACTION

- Using HMM Profiles we extracted HMM matrix for every protein sequence.
- Size of each profile is (L,20) where L is the length of the sequence.
- Amino acid interactions are captured using the proposed SXGbg technique. Features are extracted from evolutionary-based profiles using this equation.

$$SXGbg(i, j) = \sum_{l=1, X \in \{0-6\}}^{L-X-1} EP_{(l, i)} \times EP_{(l+X+1, j)}$$

FEATURE EXTRACTION

We Extracted two matrix of size (20,20)

1. Normal SXG matrix using the above mentioned formula.
2. Another one is by standardizing the SXG matrix.

We passed both these matrix seperately and then copared their accuracies.

MODEL TRAINING

Deep Neural Network

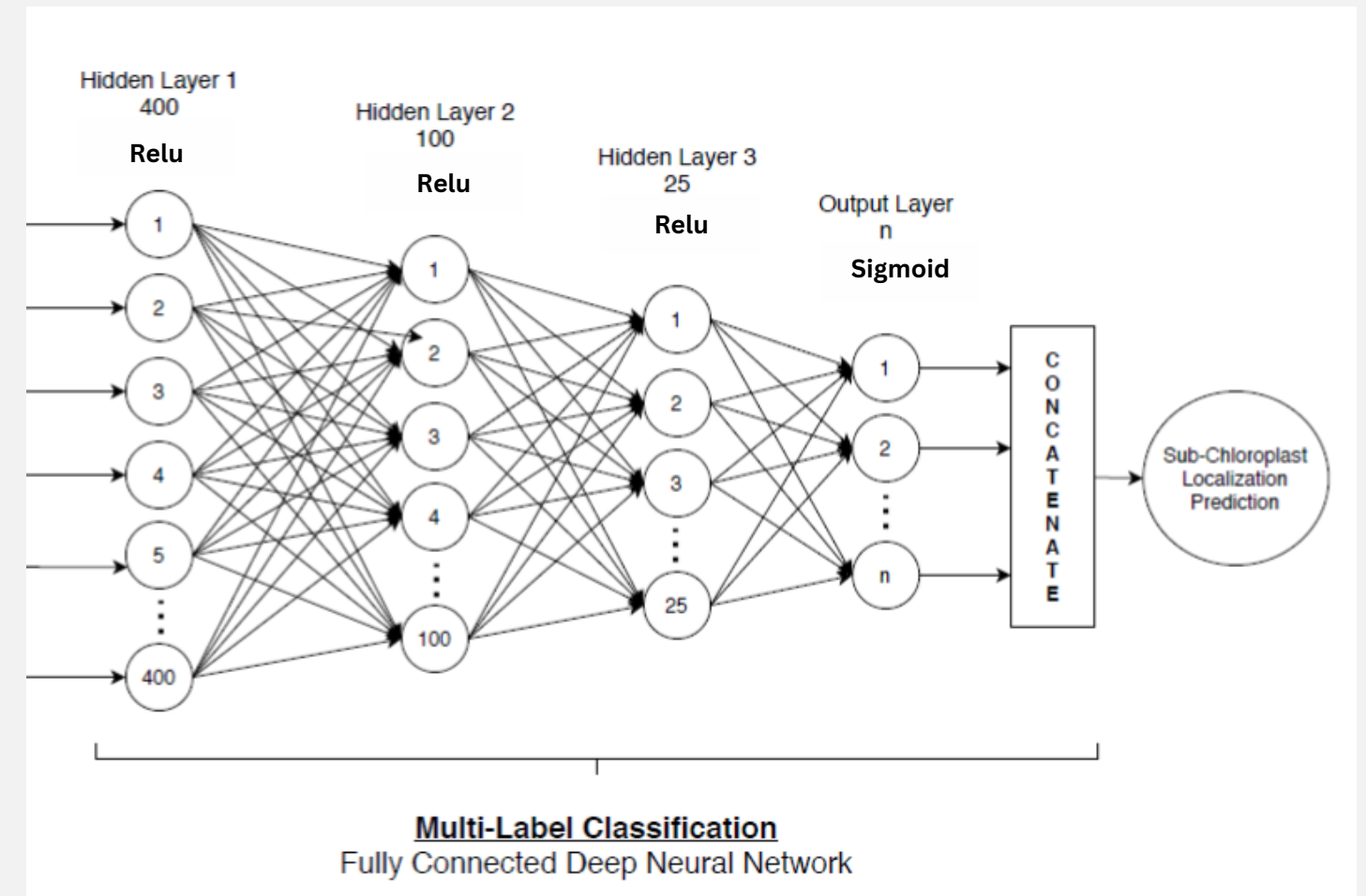
Input Shape = (batch_size, 400)

Kernel_INITIALIZER (Hidden Layers) = He_Normal

Epochs = 30 (Benchmark), 15 (Novel)

Batch_Size = 4 (Benchmark), 2 (Novel)

Callbacks = [Early Stopping, LR Scheduler]



Optimizer = Adam

Loss = Binary_Crossentropy

Metrics = Accuracy

K-FOLD CROSS VALIDATION

Crucial Technique for assessing and validating how well the model generalizes to unseen data.

Number of Splits = 5 equally sized folds

USING SXG MATRIX - HMM_BENCHMARK

	S0X	S1X	S2X	S3X	S4X	S5X	S6X
Accuracy	0.906	0.92	0.935	0.916	0.905	0.942	0.909
Jaccard Accuracy	0.645	0.72	0.78	0.733	0.708	0.818	0.669
F1	0.724	0.777	0.832	0.774	0.749	0.853	0.745
Precision	0.837	0.859	0.911	0.837	0.811	0.911	0.857
Recall	0.638	0.708	0.765	0.719	0.696	0.802	0.659
OVA	0.625	0.701	0.767	0.717	0.691	0.803	0.651
Grand Mean	0.729	0.781	0.832	0.783	0.76	0.855	0.748

STANDARDIZING SXG MATRIX - HMM_BENCHMARK

	S0X	S1X	S2X	S3X	S4X	S5X	S6X
Accuracy	0.954	0.935	0.953	0.937	0.949	0.963	0.958
Normalized Accuracy	0.865	0.807	0.854	0.787	0.845	0.891	0.876
F1	0.888	0.841	0.885	0.839	0.873	0.91	0.898
Precision	0.921	0.866	0.915	0.913	0.912	0.938	0.922
Recall	0.857	0.818	0.857	0.777	0.838	0.883	0.875
OAA	0.85	0.775	0.836	0.768	0.829	0.874	0.857
Grand Mean	0.889	0.841	0.883	0.837	0.874	0.91	0.898

USING SKIP5GRAM FEATURES - HMM_BENCHMARK

Models	Accuracy	Jaccard Accuracy	Precision	Recall	F1	OVA	Grand Mean
XGBOOST	0.88	0.57	0.79	0.57	0.67	0.55	0.67
KNN	0.85	0.54	0.71	0.53	0.61	0.52	0.62
Deep Neural Network	0.963	0.891	0.94	0.94	0.91	0.91	0.93
LSTM	0.95	0.84	0.9	0.85	0.87	0.82	0.87


For S5G Feature Modelling, Deep Neural Network achieved best results when compared with other models.

USING SXG MATRIX - PSSM_BENCHMARK

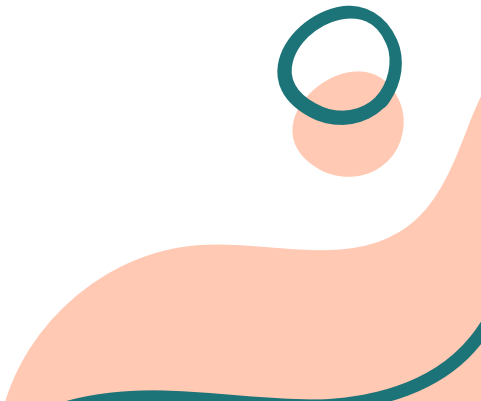
	S0X	S1X	S2X	S3X	S4X	S5X	S6X
Accuracy	0.949	0.946	0.951	0.946	0.905	0.946	0.905
Jaccard Accuracy	0.847	0.82	0.863	0.823	0.851	0.836	0.683
F1	0.877	0.857	0.887	0.864	0.878	0.87	0.716
Precision	0.909	0.908	0.916	0.911	0.916	0.903	0.765
Recall	0.848	0.812	0.86	0.822	0.844	0.839	0.673
OVA	0.824	0.803	0.836	0.801	0.829	0.813	0.668
Grand Mean	0.876	0.858	0.885	0.861	0.878	0.868	0.735

STANDARDIZING SXG MATRIX - PSSM_BENCHMARK

	S0X	S1X	S2X	S3X	S4X	S5X	S6X
Accuracy	0.971	0.97	0.955	0.96	0.955	0.949	0.968
Normalized Accuracy	0.918	0.913	0.867	0.88	0.864	0.834	0.908
F1	0.928	0.928	0.894	0.902	0.891	0.872	0.922
Precision	0.939	0.944	0.919	0.93	0.918	0.916	0.932
Recall	0.918	0.913	0.87	0.875	0.867	0.832	0.913
OAA	0.905	0.9	0.848	0.86	0.845	0.817	0.891
Grand Mean	0.93	0.928	0.892	0.901	0.89	0.87	0.922



Standardizing the obtained SXGbg Matrix resulted
in improved results when compared to normal
SXGbg matrix.



STANDARDIZING SXG MATRIX - HMM_NOVEL

	S0X	S1X	S2X	S3X	S4X	S5X	S6X
Accuracy	0.959	0.914	0.979	0.96	0.964	0.961	0.948
Jaccard Accuracy	0.885	0.715	0.945	0.89	0.909	0.87	0.869
F1	0.903	0.781	0.945	0.902	0.912	0.899	0.869
Precision	0.916	0.874	0.945	0.914	0.933	0.942	0.893
Recall	0.89	0.706	0.945	0.891	0.892	0.859	0.846
OAA	0.845	0.678	0.935	0.862	0.877	0.853	0.852
Grand Mean	0.9	0.778	0.949	0.903	0.914	0.897	0.88

USING SKIP2GRAM FEATURES - HMM_NOVEL

Models	Accuracy	Jaccard Accuracy	Precision	Recall	F1	OVA	Grand Mean
XGBOOST	0.85	0.56	0.68	0.57	0.62	0.51	0.63
KNN	0.83	0.53	0.64	0.53	0.58	0.49	0.6
Deep Neural Network	0.98	0.945	0.94	0.94	0.94	0.93	0.95
LSTM	0.95	0.85	0.9	0.85	0.88	0.83	0.88

For S2G Feature Modelling, Deep Neural Network achieved best results when compared with other models.

STANDARDIZING SXG MATRIX - PSSM_NOVEL

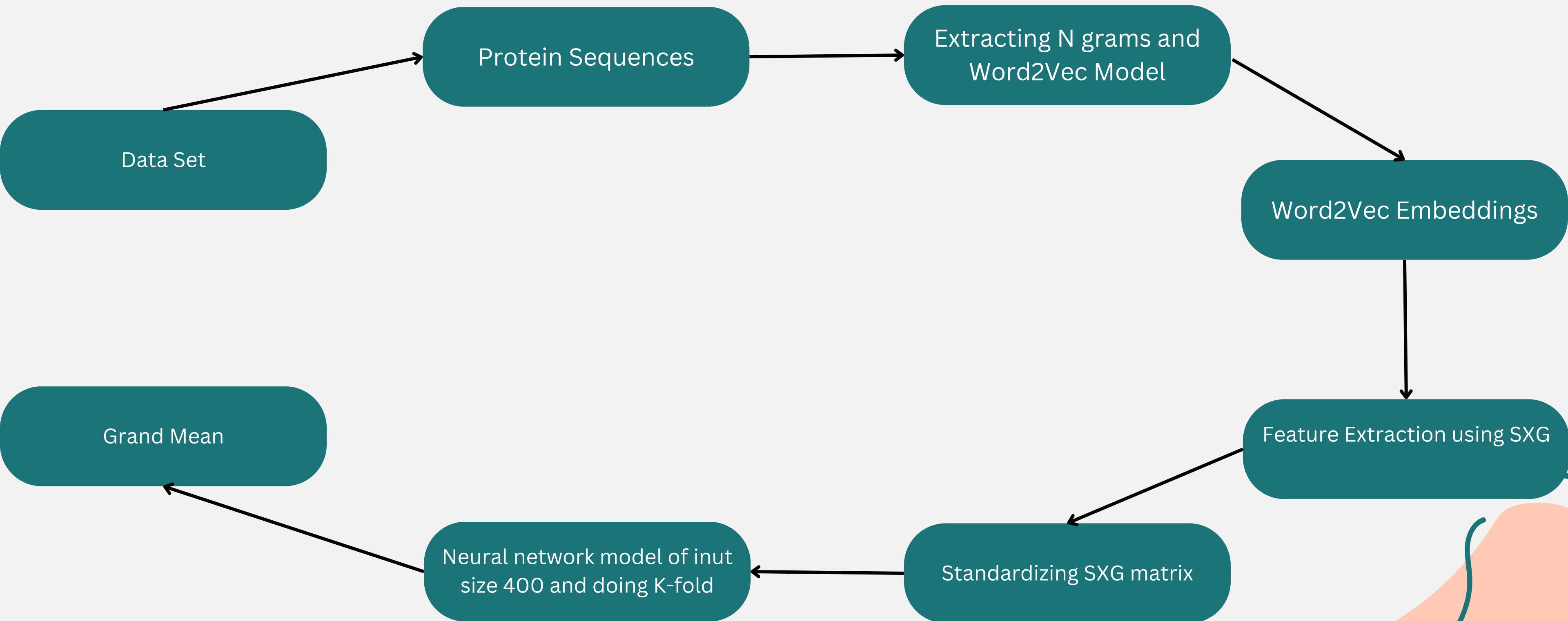
	S0X	S1X	S2X	S3X	S4X	S5X	S6X
Accuracy	0.902	0.966	0.972	0.984	0.968	0.966	0.953
Normalized Accuracy	0.931	0.898	0.918	0.96	0.915	0.915	0.882
F1	0.94	0.911	0.938	0.962	0.918	0.911	0.887
Precision	0.94	0.923	0.962	0.97	0.922	0.923	0.89
Recall	0.938	0.9	0.915	0.953	0.915	0.9	0.885
OAA	0.902	0.869	0.885	0.951	0.877	0.886	0.845
Grand Mean	0.938	0.911	0.932	0.963	0.919	0.917	0.89



MODEL 2

- Passing Word2Vec Embeddings as input to SXG for feature extraction and passing them through Deep Neural Network.

FLOW CHART



EVALUATION METRICS

- Overall Validation Accuracy
- Accuracy
- Jaccard Accuracy
- Precision
- Recall
- F1-score

GRAND MEAN = (ACC + VAL ACC + JAC ACC + PRECISION + RECALL + F1)/6

FEATURE EXTRACTION

N-GRAM WORD2VEC EMBEDDINGS

BIGRAMS

Protein Sequence: MLDLC

Bi-Grams: ML, LD, DL, LC.



WORD2VEC MODEL

Input: Bi-Grams

Vector Size: 20

Window: 7



WORD2VEC EMBEDDINGS

Size: $(L-N+1, 20)$

L is the length of sequence

N is N-Gram Model

```
2grams \
0 [MD, DL, LC, CS, SS, ST, TG, GR, RG, GA, AC, C...
1 [MD, DT, TV, VL, LM, MA, AT, TT, TP, PP, PI, I...
2 [MS, SS, SS, SS, SL, LV, VT, TS, SL, LL, LF, F...
3 [MA, AC, CR, RF, FP, PL, LH, HS, SS, SS, SP, P...
4 [MQ, QS, SA, AM, MA, AL, LS, SF, FS, SQ, QT, T...
..
116 [MA, AV, VL, LS, ST, TI, IY, YS, SI, IT, TR, R...
117 [MM, MS, SI, IP, PM, ME, EL, LM, MS, SI, IR, R...
118 [MG, GF, FL, LV, VA, AV, VM, MN, NF, FS, SP, P...
119 [MA, AY, YS, SL, LP, PT, TF, FP, PQ, QA, AL, L...
120 [MA, AA, AS, SL, LQ, QS, SA, AN, NP, PT, TL, L...

2gramembeddings
0 [[0.8774682283401489, 1.305655837059021, 2.116...
1 [[0.8774682283401489, 1.305655837059021, 2.116...
2 [[0.18904219567775726, 2.024362564086914, -1.9...
3 [[1.4109801054000854, 4.352104663848877, -1.39...
4 [[-0.3454149663448334, 0.2809176743030548, 1.5...
..
116 [[1.4109801054000854, 4.352104663848877, -1.39...
117 [[0.28322750329971313, 1.3734500408172607, -0....
118 [[0.22712849080562592, -0.39139074087142944, -...
119 [[1.4109801054000854, 4.352104663848877, -1.39...
120 [[1.4109801054000854, 4.352104663848877, -1.39...

[121 rows x 2 columns]
```


FEATURE EXTRACTION

- Using N-Gram Word2Vec Embeddings, we extracted embeddings for Bi-gram, Tri-gram and 4-gram.
- These embeddings were passed to SXG for feature extraction
- This SXG matrix is standardized as we have seen in the above model that standardized matrix has given better accuracies.

$$SXGbg(i, j) = \sum_{l=1, X \in \{0-6\}}^{L-X-1} EP_{(l, i)} \times EP_{(l+X+1, j)}$$

- This provides a high level representation of input sequence.

K-FOLD CROSS VALIDATION

Crucial Technique for assessing and validating how well the model generalizes to unseen data.

Number of Splits = 5 equally sized folds

EARLY STOPPING AND LR SCHEDULER

```
monitor = 'val_accuracy'
```

```
min_delta = 0.00005
```

```
patience = 5
```

```
verbose = 1
```

```
restore_best_weights = True
```

```
monitor= 'val_accuracy',
```

```
factor = 0.5
```

```
patience = 5
```

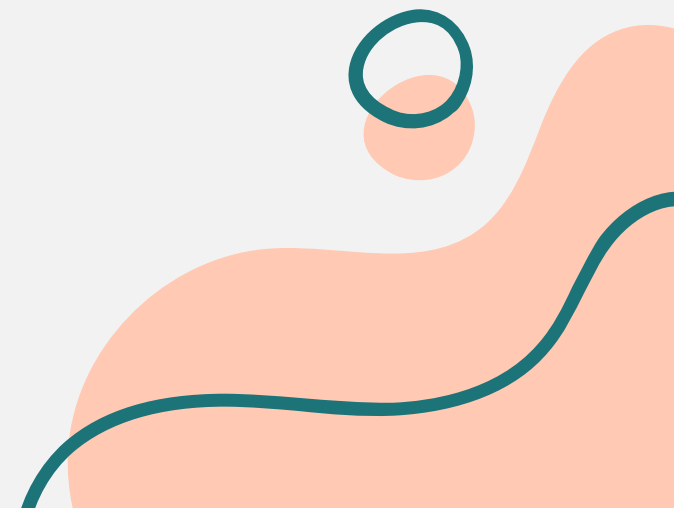
```
min_lr = 1e-6
```

```
verbose=1
```



MODEL TRAINING

1. XGBOOST Classifier
2. KNN Classifier
3. Artificial Neural Network
4. RNN and LSTM's



MODEL TRAINING

Deep Neural Network

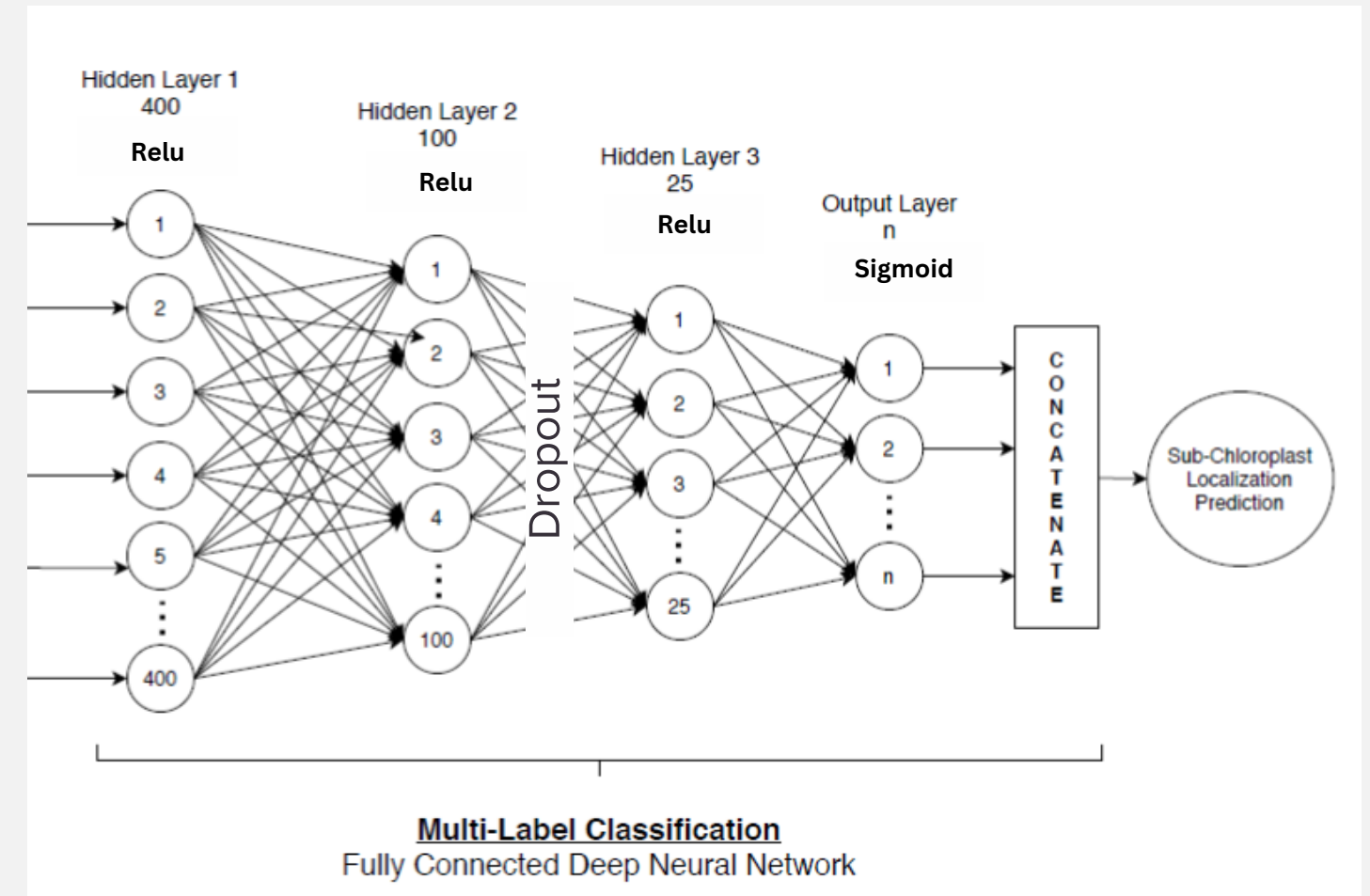
Input Shape = (batch size, 400)

Kernel Initializer (Hidden Layers) = He_Normal

Epochs = 30 (Benchmark), 15 (Novel)

Batch_Size = 4 (BenchMark), 2 (Novel)

Callbacks = [Early Stopping, LR Scheduler]



Optimizer = Adam

Loss = Binary_Crossentropy

Metrics = Accuracy

N-GRAM WORD2VEC SXG MATRIX - BENCHMARK DATASET

	BI-Gram					Tri-Gram			Four-Gram	
	S0G	S1G	S2G	S3G	S4G	S0G	S1G	S2G	S0G	S1G
Accuracy	0.952	0.956	0.952	0.95	0.93	0.94	0.95	0.96	0.79	0.79
Jaccard Accuracy	0.86	0.87	0.85	0.84	0.75	0.81	0.86	0.89	0.01	0.02
F1	0.88	0.88	0.87	0.86	0.8	0.84	0.88	0.90	0.03	0.03
Precision	0.91	0.89	0.88	0.88	0.84	0.9	0.89	0.92	0.27	0.33
Recall	0.85	0.87	0.85	0.84	0.76	0.79	0.88	0.89	0.01	0.01
OVA	0.88	0.86	0.83	0.82	0.72	0.79	0.85	0.87	0.01	0.02
Grand Mean	0.88	0.89	0.874	0.86	0.79	0.84	0.89	0.91	0.19	0.2

Best Features for Novel Dataset:

Tri-Gram + S2G
feature Modelling

TRI-GRAM WORD2VEC + S2G MAT - BENCHMARK DATASET

Models	Accuracy	Jaccard Accuracy	Precision	Recall	F1	OVA	Grand Mean
XGBOOST	0.81	0.37	0.56	0.33	0.42	0.3	0.47
KNN	0.81	0.41	0.52	0.41	0.46	0.41	0.5
Deep Neural Network	0.97	0.92	0.93	0.94	0.91	0.91	0.93
LSTM	0.92	0.73	0.81	0.74	0.77	0.71	0.78

For TRI-GRAM WORD2VEC WITH S2G Feature Modelling, Deep Neural Network achieved best results when compared with other models.

N-GRAM WORD2VEC SXG MATRIX - NOVEL DATASET

	Bi-Gram					Tri-Gram			Four-Gram	
	S0G	S1G	S2G	S3G	S4G	S0G	S1G	S2G	S0G	S1G
Accuracy	0.91	0.95	0.97	0.96	0.97	0.82	0.82	0.83	0.88	0.92
Jaccard Accuracy	0.65	0.81	0.88	0.86	0.91	0.3	0.32	0.45	0.68	0.78
F1	0.7	0.84	0.91	0.87	0.9	0.33	0.4	0.47	0.69	0.77
Precision	0.8	0.85	0.92	0.9	0.87	0.4	0.53	0.53	0.75	0.77
Recall	0.62	0.83	0.9	0.85	0.92	0.28	0.32	0.43	0.64	0.76
OVA	0.62	0.79	0.86	0.84	0.9	0.3	0.3	0.44	0.65	0.74
Grand Mean	0.72	0.84	0.91	0.88	0.91	0.41	0.45	0.52	0.72	0.79

Best Features for Novel Dataset:

Bi-Gram + S2G
feature Modelling

BI-GRAM WORD2VEC + S2G MAT - NOVEL DATASET

Models	Accuracy	Jaccard Accuracy	Precision	Recall	F1	OVA	Grand Mean
XGBOOST	0.87	0.62	0.52	0.44	0.47	0.35	0.55
KNN	0.81	0.45	0.58	0.46	0.51	0.41	0.54
Deep Neural Network	0.97	0.88	0.92	0.9	0.91	0.86	0.91
LSTM	0.95	0.84	0.9	0.85	0.87	0.81	0.87

For BI-GRAM WORD2VEC WITH S2G Feature Modelling, Deep Neural Network achieved best results when compared with other models.

CONCLUSION

1. Model 1 and Model 2 are performing equally well for both Benchmark and Novel datasets.
2. Combining N-Gram Word2Vec Model with SXGbg model for feature extraction provided high-level representation of protein sequences.
3. Deep Neural Network performed well compared to LSTM's and traditional ML models.

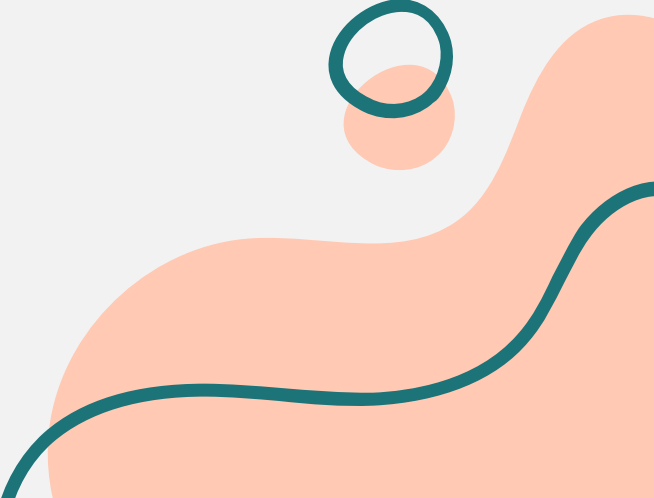


NOVELTY

1. This approach doesn't require HMM and PSSM evolutionary profiles for Feature Extraction.
2. Standardizing the SXG matrices giving better Accuracies.
3. Here we considered Jaccard Accuracy as an Evaluation Metric.



FUTURE WORK

1. Convert Protein Sequences into Graph data (Connection between nodes using Edges) and train Graph Convolutional Network (GCN) to classify based on protein subcellular localization.
 2. Use of LLM (Large Language Models) based embeddings such as GPT3 and BERT embeddings for feature extraction.
- 

INDIVIDUAL CONTRIBUTIONS

1. Gagan Vadlamudi (20bds019) and Polisetti Likhith Sai (20bds039)
 - a. Feature Extraction (SXGbg modelling, Word2Vec Model, N-Gram Word2Vec Embeddings, Standardization)
 - b. Evaluation Metrics for Model Evaluation
2. Aravind Gangavarapu and Peddisetty Venkata Sai Pranay
 - a. Model Training (KNN Classifier, XGBoost, Deep Neural Networks, LSTM's), K-Fold Cross Validation and Callbacks.
 - b. Hyperparameter Tuning

THANK
YOU

