

TABLE OF CONTENT

CERTIFICATE.....	iii
DECLARATION BY THE CANDIDATE	v
COPYRIGHT TRANSFER CERTIFICATE.....	vii
ACKNOWLEDGEMENTS.....	ix
Table of Content	xi
LIST OF FIGURES	xv
LIST OF TABLES.....	xvii
LIST OF ABBREVIATIONS.....	xix
LIST OF SYMBOLS	xxi
PREFACE.....	xxiii
Chapter 1: Introduction.....	1
1.1 Background	1
1.2 Motivation	3
1.3 Problem Statement	4
1.4 Research Objectives	5
1.5 Research Contributions	6
1.6 Outline of the Thesis	7
Chapter 2: Theoretical Background and Literature Review.....	9
2.1 Protein Secondary Structure Prediction	9
2.2 Deep Neural Network	12
2.2.1 Recurrent Neural Network	12
2.2.2 Gated Recurrent Unit.....	14
2.2.3 Long Short Term Memory.....	15
2.3 Research gaps and Findings:.....	17
2.4 Benchmark Datasets:.....	18
2.5 Performance Metrics:	18
2.6 Conclusions.....	20
Chapter 3: Protein Secondary Structure Prediction using Sequence Embedding and Bi-LSTM.....	21
3.1 Introduction	21

3.2 Protein Sequence Representation	25
3.3 Dataset Description.....	26
3.4 Model Architecture.....	27
3.4.1 Feature Extraction.....	27
3.4.2 Deep Network: Bidirectional Long Short Term Memory.....	27
3.4.3 Fully Connected Layer.....	29
3.5 Experimental Analysis.....	30
3.6 Results and Discussion	31
3.7 Conclusion.....	35
Chapter 4: Protein Secondary Structure Prediction using Character n-gram Embedding and Bi-LSTM	36
4.1 Introduction	36
4.2 Dataset Description.....	41
4.3 Problem Statement:.....	41
4.3.1 Input features	41
4.3.2 Outputs:.....	42
4.4 Methodology.....	43
4.4.1 Feature Extraction.....	43
4.4.2 Deep Network: Bidirectional LSTM	43
4.4.3 Fully Connected Layer.....	45
4.5 Experimental Details	45
4.5.1 Experimental Setup.....	45
4.5.2 Hyper-parameters.....	46
4.5.3 Training and Test Strategy.....	46
4.6 Results and Discussion	47
4.7 Conclusion.....	49
Chapter 5: Variable Length Character N-Gram Embedding Of Protein Sequences for Secondary Structure Prediction.....	50
5.1 Introduction	50
5.2 Protein Sequence Representation Method.....	53
5.3 Dataset Description.....	54
5.4 Methodology.....	55

5.4.1 Model Details	55
5.4.2 Character N-Gram	56
5.4.3 Deep Network: Bidirectional LSTM	57
5.4.4 Fully Connected Layer	57
5.5 Experimental Details	58
5.6 Results and Discussion.....	59
5.7 Conclusion.....	64
Chapter 6: Sequence-To-Sequence Modeling for Protein Secondary Structure Prediction	65
6.1 Introduction	65
6.2 Datasets	68
6.3 Proposed Model	72
6.4 Results and Discussion.....	74
6.5 Conclusion.....	76
Chapter 7: Conclusion and Future Work.....	79
7.1 Conclusions	79
7.2 Future Work	82
List of Papers Published /Communicated.....	95

LIST OF FIGURES

Figure 2.1: Recurrent Neural Network	13
Figure 2.2: Gated Recurrent Unit	14
Figure 2.3: Long Short Term Memory Network	16
Figure 3.1: Model Architecture of Protein Secondary Structure Prediction.....	29
Figure 3.2: Plot of Training Accuracy and testing Accuracy over Epochs for CullPDB32	
Figure 3.3: Plot of Training Accuracy and testing Accuracy over Epochs for SS.txt....	32
Figure 3.4: plot for Training and Testing Loss over Epochs for SS.txt.....	33
Figure 3.5: plot for Training and Testing Loss over Epochs for CullPDB.....	33
Figure 4.1: Model Architecture of Protein Secondary Structure Prediction.....	44
Figure 4.2: Plot for Training and Testing Accuracy over Epochs for ss.txt	47
Figure 4.3: Plot for Training and Testing Loss over Epochs for ss.txt.....	47
Figure 5.1: Proposed Model for Protein Secondary Structure Prediction	56
Figure 5.2: Training and Testing Loss over Epochs for ss.txt.....	60
Figure 5.3: Training and Testing Accuracy over Epochs for ss.txt	60
Figure: 6.1 Sequence to Sequence using LSTM [8]	69
Figure 6.2: Sequence to sequence model for protein secondary structure prediction	72

LIST OF TABLES

Table 3.1: DSSP 8-class to 3-class Conversion	26
Table 3.2: Configuration of Training Hyper-parameters	31
Table 3.3: Performance of Bi-RNN, Bi-GRU and Bi-LSTM on ss.txt dataset	32
Table 3.4: Comparison of the performance of various	34
Table 3.5: Performance Accuracy comparison with other Methods on Dataset RS126.	35
Table 4.1: DSSP 8-class to 3-class conversion.....	42
Table 4.2: Configuration of Training Hyper-parameters	46
Table 4.3: Performance Q3 Accuracy comparison with other methods on CASp9 dataset	48
Table 4.4: Performance Q3 Accuracy comparison with other methods on RS126 dataset.	49
Table 5.1: Hyper-parameters Value.....	58
Table 5.2: Performance Comparison of Bi-RNN, Bi-GRU and Bi-LSTM	59
Table 5.3: Q3 accuracy for different value of n-gram	61
Table 5.4: Q3 accuracy for different combination of n-gram.....	62
Table 5.5: Q3 value comparison with other Methods on dataset CASP9.....	63
Table 5.6: Q3 value comparison with other Methods on dataset RS126.....	63
Table 6.1: Amino acid one-hot Encoding Representation.....	69
Table 6.2: Comparison of the performance of various single-sequence based prediction on Cullpdb.....	75
Table 6.3: Comparison of various single-sequence based prediction on data1199	76