

Artificial Intelligence-based Novel Techniques for Accelerating Drug Discovery



Thesis submitted in partial fulfillment
for the award of the degree of

Doctor of Philosophy

by

Vishakha Singh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 20071505

Year 2024

CERTIFICATE

It is certified that the work contained in the thesis titled "*Artificial intelligence-based novel techniques for accelerating drug discovery*" by *Vishakha Singh* has been carried out under my supervision, and this work has not been submitted elsewhere for a degree. It is further certified that the student has fulfilled all requirements of Comprehensive Examination, Candidacy, and SOTA for the award of Ph.D. Degree.


Prof. Sanjay Kumar Singh
Professor

Dept. of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi

पर्यवेक्षक/Supervisor
संगणक विज्ञान एवं अभियंत्रिकी विभाग
Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी/Varanasi-221005

DECLARATION BY THE CANDIDATE

I, *Vishakha Singh*, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of *Prof. Sanjay Kumar Singh* from *December 2020* to *January 2024* at *Department of Computer Science and Engineering*, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not wilfully copied any other's work, paragraphs, text, data, results, *etc.* reported in journals, books, magazines, reports, dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my own work.

Date: 25.07.24

Place: Varanasi



Vishakha Singh

CERTIFICATE BY THE SUPERVISOR

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.



Prof. Sanjay Kumar Singh

Professor

Dept. of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi



Signature of Head of the Department

आचार्य एवं विभागाध्यक्ष
Professor & Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(बनारस हिन्दू यूनिवर्सिटी)
(Banaras Hindu University)
वाराणसी-221005

सुपरविज़र
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी/Varanasi-221005

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Artificial intelligence-based novel techniques for accelerating drug discovery

Name of the Student: Vishakha Singh

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy*.

Date: 25.07.24

Place: Varanasi



Vishakha Singh

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Dedicated to my beloved parents.

ACKNOWLEDGEMENT

The completion of my doctorate as a Prime Minister Research Fellow has provided me with a wonderful opportunity to express my heartfelt gratitude towards people who were an integral part of this journey. I am grateful to everyone who helped make this dissertation possible by constantly motivating and assisting me to the best of their efforts and resources. First and foremost, I would like to thank my esteemed supervisor, Prof. Sanjay Kumar Singh, for his extraordinary support and supervision during my Ph.D. He guided and motivated me using his knowledge and expertise, from the beginning until the end of my degree, by nudging me in the right direction. I am also grateful to my doctoral committee (Prof. Neeraj Sharma and Dr. Ravi Shankar Singh) for reviewing my research progress and providing invaluable insights as to how I can improve upon it. This acknowledgment cannot be complete without my parents and siblings, to whom I owe this thesis. I want to express my heartfelt gratitude to my mother, Dr. Punam Singh, and father, Mr. Srinivas Singh, for their unwavering support and understanding and for being my pillars of strength throughout this journey. I am also thankful to my sister, Dr. Aakansha Singh, and brother, Mr. Shashank Shekhar Singh, for motivating and encouraging me. As I complete my doctorate, my heart goes out to my late grandmother, Mrs. Sushila Singh, who was a crucial part of my journey. This note of thanks would be incomplete without mentioning my friends and colleagues (Ms. Deepti Chauhan, Dr. Ritesh Sharma, Dr. Ekta Sharma, Mr. Jayashankara M. and Mr. Rajat Gupta), to whom I am grateful for their encouragement and constructive criticism, which enriched the quality of this thesis. Lastly, I am obliged to all the department's technical and non-technical staff members, Mr. Ravi Kumar Bharti, Mr. Shubham Pandey, and Mr. Prakhar Kumar.

This thesis marks a significant milestone in my career. With heartfelt gratitude, I look forward to apply the lessons learned from this experience to my future endeavors.

Vishakha Singh
(Vishakha Singh)

Contents

List of Figures	v
List of Tables	viii
List of Abbreviations	ix
List of Symbols	xi
Preface	xiii
1 Introduction	1
1.1 Problem definition	1
1.2 Existing works	2
1.3 Motivation	4
1.4 Preliminaries	6
1.4.1 Bidirectional long short-term memory networks (biLSTM)	6
1.4.2 Temporal Convolutional Networks (TCNs)	6
1.4.3 Bidirectional encoding representations from transformers (BERT)	7
1.4.4 Non-dominated sorting genetic algorithms (NSGA)-II	8
1.4.5 Gravitational Search Algorithm (GSA)	8
1.4.6 Continual Learning	8
1.4.7 Transfer Learning	9
1.5 Contributions	10
2 Using the stacked ensemble technique for the prediction and discovery of antibacterial peptides	13
2.1 Introduction	14
2.2 Data and preliminaries	17
2.2.1 Dataset	17

2.2.2	Long Short-Term Memory Networks	17
2.2.3	Attention mechanism	19
2.3	Proposed work	19
2.3.1	The base level	21
2.3.2	The meta level	21
2.4	Experiments, results, and discussions	21
2.4.1	Evaluation Criteria	22
2.4.2	Performance Evaluation	22
2.4.3	Discovering ABPs in proteins	26
2.5	Conclusion	28
3	A multi-layered continual-learning based architecture for discovering antibacterial peptides	29
3.1	Introduction	30
3.2	Data and preliminaries	34
3.2.1	Dataset	34
3.2.2	Continual Learning	34
3.2.3	Temporal Convolutional Networks (TCNs)	36
3.3	Proposed Work	38
3.3.1	MSTCN-ABPpred (BL) model	38
3.3.2	MSTCN-ABPpred (CL) model	39
3.4	Experiments, Results, and Discussions	44
3.4.1	Evaluation Criteria	45
3.4.2	Performance Evaluation	45
3.5	Conclusion	54
4	A resource-efficient deep learning model for the discovery of antiviral peptides	57
4.1	Introduction	58
4.2	Data and preliminaries	63
4.2.1	Dataset	63
4.2.2	Word embeddings	63
4.2.3	Temporal Convolutional Networks	64
4.2.4	Depth-wise separable convolutions	64
4.3	Proposed Work	65
4.4	Experiments, Results, and Discussions	67
4.4.1	Evaluation Criteria	67

4.4.2	Performance Evaluation	68
4.4.3	Predicting AVPs using the web application	69
4.5	Conclusion	72
5	Using genetic algorithms and explainable AI for classifying and optimizing neurological peptides	75
5.1	Introduction	76
5.2	Data and preliminaries	80
5.2.1	Dataset	80
5.2.2	Non-dominated sorting genetic algorithm-II	81
5.2.3	BERT	82
5.2.4	Needleman Wunsch Algorithm	83
5.2.5	Captum	83
5.3	Proposed work	83
5.3.1	Phase 1: The BERT-NeuroPred model	84
5.3.2	Phase-2: The NSGA-NeuroPred framework	85
5.4	Experiments, results, and discussions	93
5.4.1	Evaluation Criteria	93
5.4.2	Analysis of phase-1	94
5.4.3	Analysis of phase-2	101
5.5	Conclusion	102
6	Optimizing blood-brain barrier penetrating peptides using explainable AI	105
6.1	Introduction	106
6.2	Data and preliminaries	108
6.2.1	Dataset	108
6.2.2	Gravitational search algorithm (GSA)	109
6.3	Proposed work	111
6.3.1	Phase-1: Construction of ML-B3P2pred and DL-B3P2pred	111
6.3.2	Phase-2: The Hybridised Gravitational Search Algorithm (HyGSA)	112
6.4	Experiments, results, and discussions	126
6.4.1	Experiments conducted in phase-1	127
6.4.2	Using the HyGSA to find novel B3P2s	132
6.5	Conclusion	133

7 Conclusion and Future Directions	135
7.1 Conclusion	135
7.2 Future Directions	136
List of Publications	139
Bibliography	141

List of Figures

1.1	Therapeutic peptide-based drug discovery	5
1.2	Working of a biLSTM cell	6
1.3	A TCN block	7
1.4	Layout of the thesis	10
2.1	The architecture of the StaBle-ABPpred model	20
2.2	Confusion matrices for various models on test set	23
3.1	Architecture of MSTCN-ABPpred model	39
3.2	The working of the continual learning module. Step 1 is performed only once on the initial training set to build the MSTCN-ABPpred (BL) model. After that, steps 2-8 are repeated for further re-training (Steps 5 and 6 are not executed in case of re-training using approaches 1 and 3). To predict peptides in step 3 from the protein entered in step 2, the latest model (which has been re-trained on all the protein sequences entered before the current one) is always used.	41
3.3	Box and whisker plots depicting performance of different models on test set based on accuracy(%), f1-score(%), and AUC(%)	48
3.4	Performance of MSTCN-ABPpred (BL), MSTCN-ABPpred (A1), MSTCN-ABPpred (A2), MSTCN-ABPpred (A3), and MSTCN-ABPpred (A4) on an independent set of (a) ABPs active against ESKAPEE pathogens, and (b) general non-ABPs	49
3.5	Performances (on the test set (Te)) of MSTCN-ABPpred (BL), MSTCN-ABPpred (A1), MSTCN-ABPpred (A2), MSTCN-ABPpred (A3) and MSTCN-ABPpred (A4) that were re-trained using approaches 1,2,3 and 4, respectively on (a) 7 sequences, (b) 14 sequences, (c) 21 sequences, (d) 28 sequences, (e) 35 sequences, (f) 42 sequences, (g) 49 sequences, (h) 56 sequences, (i) 63 sequences	50

3.6	Visualization of peptides in the discovered set (D), and the initial training set (T) using the isomap technique.	51
4.1	The Deep-AVPiden (DS) architecture	66
4.2	Confusion matrices obtained for various models including Deep-AVPiden on the test set	69
4.3	Alpha-helical representations of AVPs discovered in the plant, mammal, and fish proteins	70
4.4	Scatter plot showing the distribution of AVPs predicted in the plant, mammal, and fish antiviral proteins, along with the AVPs and non-AVPs in the training set	71
5.1	Architecture of NPpred framework	86
5.2	Performance of different deep learning models on the test set	96
5.3	Performance of different stacked models on the test set	96
5.4	Confusion matrix depicting performance of the proposed and state-of-the-art models on the test set	102
5.5	Isometric mapping of peptides predicted by NPpred	103
6.1	The overall process involved in building HyGSA	113
6.2	Architecture of ML-B3P2pred and DL-B3P2pred models	114
6.3	Particle encoding	114
6.4	The important features and their weightage as per (a) SHAP, (b) ELI5, and (c) Yellowbrick (Shapiro-Wilk algorithm)	126
6.5	Relationship between feature and SHAP values. A single dot represents one data point.	127
6.6	Relationship between the feature and the Shapley values (a) pI, (b) AAC_Y, (c) normalized mw, (d) AAC_K, (e) AAC_C, (f) instability index, (g) AAC_W, (h) AAC_V, (i) AAC_F	128
6.7	The flowchart depicting the working of HyGSA on a dataset comprising cell-penetrating peptides (CPPs)	130
6.8	Isometric mapping of the B3P2s discovered by HyGSA, along with other CPPs and the peptides present in our dataset	131
6.9	(a)-(d) show the scatter plot of objective values and the fitness values of all the particles. (e)-(j) show the pairwise relation between the values of objectives obtained by different particles	131

List of Tables

2.1	Performance of various models on the test set	23
2.2	ANOVA on accuracy (%) of various models	24
2.3	ANOVA on precision (%) of various models	25
2.4	ANOVA on f1-score (%) of various models	25
2.5	Top fifteen ABPs as per the similarity score with annotated AMPs. The first row in column 7 contains the identified ABP, and the second row comprises the matching AMP found using the BLAST tool.	27
3.1	Performance of various models on the test set	46
3.2	Performance of the model on the test set after re-training on proteins .	52
3.3	ANOVA on accuracy (%) of the model after several rounds of retraining	52
4.1	Comparison of Deep-AVPiden with existing models on test set	68
4.2	The AVPs discovered in the proteins of mammals, fish, and plants, with a probability score $\geq 90\%$ and showing some sequence similarity with the AMPs existing in public databases.	70
4.3	The AVPs discovered using Deep-AVPiden (DS) were subjected to BLAST analysis. Column 5 shows the method used to validate the AMPs similar to them as antimicrobial and/or antiviral (as mentioned in column 6). Column 7 consists of the similar AAs of discovered peptides and the ones found by BLAST analysis.	71
5.1	ANOVA on accuracy (%) of all the proposed models after training on ten different dataset splits	97
5.2	ANOVA on f1-score (%) of all the proposed models after training on ten different dataset splits	97
5.3	ANOVA on AUC (%) of all the proposed models after training on ten different dataset splits	98

5.4	ANOVA on accuracy (%) of all the proposed models after training on ten different dataset splits	98
5.5	ANOVA on F1-score (%) of all the proposed models after training on ten different dataset splits	99
5.6	ANOVA on AUC (%) of all the proposed models after training on ten different dataset splits	100
5.7	Comparison of BERT-NeuroPred with state-of-the-art (SOTA) models.	101
6.1	Comparison of DL-B3P2pred with existing state-of-the-art (SOTA) models on benchmark test sets.	129
6.2	Parameters used in HyGSA	129

Abbreviations

Abbreviation	Description
ABPs	Antibacterial Peptides
AI	Artificial Intelligence
ANOVA	Analysis of Variance
AVPs	Antiviral Peptides
AUC	Area under the receiver operating characteristic curve
B3P2s	Blood-Brain Barrier Penetrating Peptides
BERT	Bidirectional Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
ESKAPEE	Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp. and Escherichia coli
GSA	Gravitational Search Algorithm
NP	Neurological Peptides
ReLU	Rectified Linear Unit
RFs	Random Forests
SHAP	Shapley Additive explanations
SVM	Support Vector Machine
TCNs	Temporal Convolutional Networks

List of Symbols

Symbol	Description
a_t	attention vector
$\alpha_{t,t'}$	attention weight
\tilde{c}_t	candidate state
c_t	cell state
cv_t	context vector
fit	fitness value
f_t	forget gate
h_t	hidden state
i_t	input gate
o_t	output gate
Pop	population of peptides
r_t	reset gate
z_t	update gate