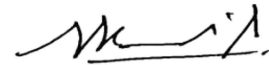


CERTIFICATE

It is certified that the work contained in the thesis titled “*Combating Antimicrobial Resistance with Artificial Intelligence*” by *Ritesh Sharma* has been carried out under my supervision and that 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 and HOD

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

DECLARATION BY THE CANDIDATE

I, **Ritesh Sharma**, 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 **July 2019** to **April 2023** 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 willfully 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: 18/09/2023

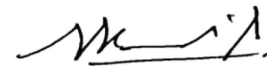
Place: Varanasi



Ritesh Sharma

CERTIFICATE BY THE SUPERVISOR

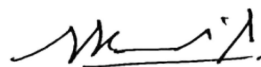
This is to certify that the above statement made by the candidate is correct to the best of my knowledge.



Prof. Sanjay Kumar Singh

Professor

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



Signature of Head of the Department

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Combating Antimicrobial Resistance with Artificial Intelligence

Name of the Student: Ritesh Sharma

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:18/09/2023

Place: Varanasi



Ritesh Sharma

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.

This thesis is dedicated to GOD and my beloved family.

ACKNOWLEDGEMENT

I bow down to the Lord Almighty, whose grace guided me from inception to the completion of this research work. Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I am grateful to everyone who helped make this dissertation possible. Firstly, I would like to thank my esteemed supervisor, Prof. Sanjay Kumar Singh, for his invaluable supervision, support, and tutelage during my Ph.D. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I hope that one day I would become as good an advisor to my students as Prof. Sanjay Kumar Singh has been to me. I am thankful to my Doctoral Committee (Prof. Neeraj Sharma and Dr. Ravi Shankar Singh) for reviewing my work and helping me with invaluable suggestions. I am grateful to my collaborators, who helped me a lot. I am also grateful to all the technical and non-technical staff members of the department (Dr. Ram Prasad Meena, Ravi Kumar Bharti, Shubham Pandey, and Prakhar Kumar) for their help and precious assistance. I also want to thank my friends (Vishakha, Jai Shankar, Priyatosh, Sushant, Sandeep, Vipin) for their continuous support throughout this journey. This acknowledgment would not be complete without mentioning the painstaking efforts and patience of my families. No words are adequate to express my indebtedness to my parents for their support, blessings, and good wishes. I owe this thesis to my parents, my in-laws, and other family members who always stood by me and provided strength in pursuing this work. I would like to express my heartfelt gratitude to my father, Capt. Y. R. Sharma & mother, Mrs. Anjana Sharma, without their unconditional support, this task would not have been accomplished. I am also grateful to my father-in-law Mr. R. P Duvedi & mother in law Mrs. Lata Duvedi, for their continuous motivation. I would like to say specially thanks to my brother (Nirdesh Kumar Sharma), sister-in-law (Shanu), brother-in-law (Abhinav) for their continuous support.

I am also very much thankful to my wife Dr. Ekta Sharma. She has been the pillar of support during this journey and without her sacrifices this thesis would not have

been completed.

Lastly, I express my hearty thanks to those I missed mentioning by name, who helped directly or indirectly and co-operated with me a lot in completing this Ph.D. research work.



Ritesh Sharma

Contents

List of Figures	xi
List of Tables	xvii
List of Abbreviations	xix
List of Symbols	xxi
Preface	xxiii
1 Introduction	1
1.1 Introduction	1
1.2 Preliminaries	4
1.2.1 Biological aspects	4
1.2.2 Computational Aspects	6
1.3 Motivation and Objective of the Thesis	14
1.4 Contributions	18
2 Accelerating the Discovery of Peptide-based Antifungal Drugs using Artificial Intelligence	21
2.1 Introduction	22
2.2 Materials and Methods	26
2.2.1 Dataset Collection	26
2.2.2 Proposed Framework	27
2.3 Experiments and Results	30
2.3.1 Experimental Configuration	31
2.3.2 Performance Metrics	31
2.3.3 Assessment Procedure	31
2.3.4 Results obtained from proposed Framework	32

2.3.5	Ablation Studies	33
2.3.6	Additional Experiments	34
2.3.7	Performance of proposed model Deep-AFPpred and other AFP prediction tools on test data	38
2.4	Prediction of AFPs in the Antifungal Proteins	40
2.5	Web Server	42
2.6	Summary	44
3	Discovery of Peptide-based Antiviral Drugs using Artificial Intelligence	47
3.1	Introduction	48
3.2	Materials and Methods	51
3.2.1	Dataset Collection	51
3.2.2	Proposed Framework	53
3.3	Experiments and Results	55
3.3.1	Experimental Configuration	56
3.3.2	Performance Metrics	56
3.3.3	Assessment Procedure	56
3.3.4	Results obtained from the proposed Framework	56
3.3.5	Additional Experiments	57
3.3.6	Performance of proposed model Deep-AVPpred and other AVP prediction tools on test data	61
3.4	Prediction of AVPs in the Antiviral Proteins	63
3.5	Web Server	65
3.6	Summary	67
4	Identifying critical amino acids in the Antibacterial molecules using Explainable Artificial Intelligence	69
4.1	Introduction	70
4.2	Materials and methods	73
4.2.1	Dataset Collection	73
4.2.2	Proposed Framework	75
4.2.3	Local interpretable model-agnostic explanations (LIME)	79
4.3	Experiments and Results	80
4.3.1	Experimental Configuration	80
4.3.2	Performance Metrics	80
4.3.3	Assessment Procedure	80

4.3.4	Result obtained from the proposed Model	81
4.3.5	Ablation Studies	81
4.3.6	Performance of the proposed model on test data	85
4.3.7	Results from various experiments conducted using LIME	85
4.4	Prediction of ABPs	90
4.5	Web Server	92
4.6	Summary	92
5	Identifying Peptide-based Antibacterial Drugs effective against ESKAPEE pathogens using Artificial Intelligence	95
5.1	Introduction	96
5.2	Materials and methods	99
5.2.1	Dataset	99
5.2.2	Proposed Framework	100
5.3	Experiments and Results	104
5.3.1	Experimental Configuration	105
5.3.2	Performance Metrics	105
5.3.3	Assessment Procedure	105
5.3.4	Results obtained by the proposed Model	106
5.3.5	Ablation Studies	106
5.3.6	Additional Experiments	112
5.3.7	Performance of the proposed model ESKAPEE-MICpred on test data	115
5.4	Prediction of MIC values against ESKAPEE	115
5.5	Web Server	117
5.6	Summary	117
6	Artificial Intelligence based Discovery of low hemolytic therapeutic peptides	119
6.1	Introduction	120
6.2	Materials and methods	124
6.2.1	Dataset	124
6.2.2	Proposed Framework	125
6.3	Experiments and Results	129
6.3.1	Experimental Configuration	129
6.3.2	Performance Metrics	129
6.3.3	Assessment Procedure	130

6.3.4	Results obtained by the proposed Framework	133
6.3.5	Ablation Studies	133
6.3.6	Additional Experiments and Analysis	140
6.4	Web server	141
6.5	Conclusion	143
7	Conclusion and Future Directions	147
7.1	Conclusion	147
7.2	Future Directions	149
	List of Publications	150
	Bibliography	151

List of Figures

1.1	Antimicrobial Resistance (Source:[1])	5
1.2	Spread of Antimicrobial resistance	5
1.3	AI in Drug Discovery (Source:[2])	14
1.4	AI in peptide based antimicrobial drug discovery.	15
1.5	Layout of the Thesis	18
2.1	Proposed Framework	27
2.2	Comparison of results obtained from the 1DCNN-BiLSTM using pretrained embeddings (PESTV) and amino acid encodings (PAM250, BLOSUM62, OHE).	34
2.3	Comparison of results obtained from the PESTV + 1DCNN-BiLSTM deep learning algorithm and HCF1 + machine learning algorithms (XGBOOST, RF, NB, SVM, LR, KNN).	36
2.4	Comparison of results obtained from the PESTV + 1DCNN-BiLSTM deep learning algorithm and HCF2 + machine learning algorithms (XGBOOST, RF, NB, SVM, LR, KNN).	37
2.5	Comparison of results obtained from the PESTV + 1DCNN-BiLSTM deep learning algorithm and HCF3 + machine learning algorithms (XGBOOST, RF, NB, SVM, LR, KNN).	37
2.6	Helical wheel representation of proposed AFPs.	40
2.7	Classify query peptide as AFP/Non-AFP.	42
2.8	Identify AFPs from protein sequence.	43
3.1	Proposed Framework.	54
3.2	Comparison of results obtained from proposed model Deep-AVPpred and machine learning based models.	59
3.3	Comparison of results obtained from proposed model Deep-AVPpred and Meta-models.	60

3.4	Comparison of results obtained from proposed model Deep-AVPpred and ANN.	60
3.5	Helical wheel representation of proposed AVPs.	65
3.6	Classify query peptide as AVP/Non-AVP	66
3.7	Identify AVPs from protein sequence	66
4.1	Proposed Framework	75
4.2	Comparison of results obtained from Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN taking one at a time.	82
4.3	Comparison of results obtained by combining any two of Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN.	83
4.4	Comparison of results obtained by combining any three of Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN.	83
4.5	Comparison of the best results obtained for all the cases.	85
4.6	Mean contribution of amino acid towards ABP class.	86
4.7	Mean contribution of amino acid towards Non-ABP class.	86
4.8	Predictions made by LIME for sample peptides from each of (A) True Positive, (B) True Negative, (C) False Positive and (D) False Negative categories	87
4.9	Conversion of ABP to Non-ABP after removing critical amino acid (s) (B) Conversion of Non-ABP to ABP after removing critical amino acid (s).	89
4.10	Identify critical amino acids in ABP.	92
5.1	Proposed Framework.	97
5.2	Number of peptides active against each bacterial species of the ESKAPEE group.	100
5.3	Distribution of ABPs (A) across a wide range of MIC values (B) across a narrow range of log (MIC) values (C) across a narrow range of log (MIC) values after standardization.	101
5.4	The variation of training and validation MSE for BiGRU, BiLSTM, BiTCN, and 1DCNN	104
5.5	Predicted versus actual activity values for peptides.	106
5.6	Comparison of results obtained by proposed ensemble classifier (ESKAPEE-MICpred), BiGRU-PESTV, BiLSTM-PESTV, BiTCN-PESTV, and 1DCNN-PESTV.	108

5.7	Comparison of results obtained by BIGRU algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).	110
5.8	Comparison of results obtained by BILSTM algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).	111
5.9	Comparison of results obtained by BITCN algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).	111
5.10	Comparison of results obtained by 1DCNN algorithm when it is utilized with pretrained embeddings (PESTV) and non-pretrained embeddings (PAM 250, BLOSUM 62, NNAA, and OHE).	112
5.11	Predicting the MIC value of antibacterial peptides against ESKAPEE pathogens.	116
6.1	Proposed Framework.	123
6.2	Comparison of results obtained on applying both HCF and DLF with (i) average combiner (ACE-HD), (ii) majority voting (MVE-HD), (iii) fuzzy non-linear (FNE-HD), (iv) fuzzy gompertz (FGE-HD), (v) fuzzy distance (FDE-HD), (vi) BiLSTM (BiLSTM-HD), (vii) BiTCN (BiTCN-HD), (viii) 1DCNN (1DCNN-HD), and (ix) Min/Max combiner (MCE-HD).	134
6.3	Comparison of results obtained on applying both HCF and DLF with (i) BiLSTM (BiLSTM-HD), (ii) BiTCN (BiTCN-HD), (iii) 1DCNN (1DCNN-HD), and (iv) Min/Max combiner (MCE-HD).	135
6.4	Comparison of results obtained on using only DLF with BiLSTM (BiLSTM-DL) and on applying both HCF and DLF with BiLSTM (BiLSTM-HD).	136
6.5	Comparison of results obtained on using only DLF with BiTCN (BiTCN-DL) and on applying both HCF and DLF with BiTCN (BiTCN-HD).	136
6.6	Comparison of results obtained on using only DLF with 1DCNN (1DCNN-DL) and on applying both HCF and DLF with 1DCNN (1DCNN-HD).	137
6.7	Comparison of results obtained on using only DLF with min/max combiner (MCE-DL) and on applying both HCF and DLF with min/max combiner (MCE-HD).	137
6.8	Comparison of results obtained on using HCF with classification module (CM-HC) and on applying both HCF and DLF with min/max combiner (MCE-HD).	140

6.9	Comparison of results obtained by meta machine learning models and our proposed framework (MCE-HD).	141
6.10	Activity prediction for query peptides.	142
6.11	Mutation Analysis by replacing amino acid present at particular location with remaining nineteen natural amino acids.	143
6.12	Residue Scan by substituting each amino acid (one at a time) present in the query peptide with the entered amino acid.	144

List of Tables

2.1	Results obtained by utilizing PESTV with a 1DCNN-BiLSTM deep learning algorithm.	32
2.2	Results obtained by utilizing amino acid encodings from PAM250, BLOSUM62, and OHE with a 1DCNN-BiLSTM deep learning algorithm.	33
2.3	Results obtained by utilizing HCF1 with various machine learning algorithms.	35
2.4	Results obtained by utilizing HCF2 with various machine learning algorithms.	35
2.5	Results obtained by utilizing HCF3 with various machine learning algorithms.	36
2.6	Results obtained from proposed model Deep-AFPpred and other AFP prediction tools	38
2.7	Proposed peptides for wet-lab synthesis and experimentation.	39
3.1	Results obtained from the proposed framework	57
3.2	Result obtained by the proposed model across five runs.	57
3.3	Results obtained by various machine learning models.	59
3.4	Results obtained by Meta models	59
3.5	Results obtained by ANN.	59
3.6	Results obtained from proposed model Deep-AVPpred and other AVP prediction tools	61
3.7	Proposed peptides for wet-lab synthesis and experimentation.	62
4.1	Result obtained from the ensemble classifier constructed by combining Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN algorithms.	81
4.2	Result obtained from Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN	82
4.3	Results obtained from the ensemble classifiers constructed by combining any two of the Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN algorithms.	82

4.4	Result obtained from the ensemble classifiers constructed by combining any three of the Bi-GRU, Bi-LSTM, Bi-TCN and 1DCNN algorithms.	83
4.5	Results obtained from the proposed model LGTC-SV on test data	84
4.6	Antibacterial peptides derived from bacteriocin of the ESKAPEE group of bacteria and proposed for wet-lab synthesis and experimentation.	91
5.1	Results obtained from proposed framework.	105
5.2	Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising PESTV	107
5.3	Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising OHE	109
5.4	Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising PAM250	109
5.5	Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising BLOSUM62	110
5.6	Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising NNAA	110
5.7	Results obtained by ANN utilising HCF	113
5.8	Results obtained by BiGRU, BiLSTM, BiTCN and 1DCNN utilising BOTH HCF AND PESTV	113
5.9	Results obtained from the proposed model ESKAPEE-MICpred on test data.	113
5.10	Antibacterial peptides identified from the proteins sequences.	114
5.11	Antibacterial peptides identified from the therapeutic peptides.	114
6.1	Criteria for classifying a peptide as high-hemolytic peptide (HEP) and low-hemolytic peptide (LEP) [3, 4, 5]	125
6.2	Results obtained by BiTCN, BiLSTM, 1DCNN, various ensemble techniques utilising HCF and DLF	131
6.3	Results obtained by Min/Max combiner algorithm utilizing HCF and DLF	131
6.4	Results obtained by BiTCN, BiLSTM, 1DCNN algorithm utilizing HCF and DLF	132
6.5	Results obtained by BiTCN, BiLSTM, 1DCNN and Min/Max combiner algorithm utilising DLF	132
6.6	Results obtained by classification module utilising HCF	139
6.7	Results obtained by utilizing different machine learning algorithms with different molecular descriptors	139

6.8 Results obtained by utilizing Meta machine learning algorithms with different molecular descriptors	139
--	-----

Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
AMR	Antimicrobial Resistance
AMPs	Antimicrobial Peptides
AVPs	Antiviral Peptides
AFPs	Antifungal Peptides
ABPs	Antibacterial Peptides
MIC	Minimum Inhibitory Concentration
WHO	World Health Organization
RBCs	Red Blood Cells
AUC	Area Under Curve
ELMo	Embeddings from Language Models
HCF	Handcrafted Features
PESTV	Pretrained Embeddings from Seq2vec
OHE	One-Hot Encoding
XGBoost	Extreme Gradient Boosting
SVM	Support Vector Machine
RF	Random Forest
LR	Logistics Regression
NB	Naive Bayes

Abbreviation	Description
KNN	K-Nearest Neighbour
ANN	Artificial Neural Network
CM	Classification Module
ERT	Extremely Randomized Trees
AB	AdaBoost
GB	Gradient Boosting
ReLU	Rectified Linear Unit
ICL	Independent Component Layer
Adam	Adaptive Moment Estimation
FDA	Food and Drug Administration
Bi-GRU	Bidirectional Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short-Term Memory
BiTCN	Bidirectional Temporal Convolutional Network
CNN	Convolutional Neural Network
LIME	Local Interpretable Model-agnostic Explanations
IC	Independent Component
GF	Genomic Features

List of Symbols

Symbol	Description
z_t	update gate
r_t	reset gate
h_t	candidate state
i_t	input gate
o_t	output gate
f_t	forget gate
L	loss function
η	learning rate
σ	activation function
x_t	embedding vector of the amino acid

Preface

The discovery of antibiotics revolutionized modern healthcare and increased the lifespan of humans by more than twenty years, but their unregulated use has resulted in antimicrobial resistance (AMR). AMR kills at least 7 lakh people in poor and middle-income countries each year, and the WHO ranks it as one of the top ten public health hazards. To counteract this condition, novel antimicrobial compounds are required, and antimicrobial peptides (AMPs) show promise in this regard. AMPs are proteins produced by diverse organisms naturally and can be classified into multiple groups, such as antiviral (AVPs), antifungal (AFPs), and antibacterial peptides (ABPs).

To avoid the high costs and time associated with identifying novel AMPs in the lab, researchers use *in-silico* tools for preliminary screening of natural sources. However, the existing tools available for this purpose have poor performance, which limits their applicability for wet-lab researchers. Thus, we have proposed AI-based frameworks for identifying AFPs, AVPs, and ABPs from natural sources in different studies.

All amino acids are not equally important in classifying a peptide. Existing AI-based tools do not provide information about the essential amino acids responsible for classifying a peptide. Therefore, we developed an explainable framework that not only classifies the peptides but also provides information about the essential amino acids responsible for classifying a peptide.

The WHO categorizes bacteria into three categories (critical, high, and medium), and the ESKAPEE pathogens are a major threat as they range from high to critical

WHO-priority pathogens. Earlier studies can identify ABPs from natural resources but do not provide minimum inhibitory concentration (MIC) values against the ESKAPEE pathogens. Thus, for identifying optimal ABPs (which work at low MIC), wet-lab researchers have to test the identified ABPs against ESKAPEE pathogens at different concentrations, leading to a loss of time and money. To address this issue, we proposed a framework that predicts the MIC values for ABPs against ESKAPEE pathogens. The proposed framework can help identify optimal ABPs that can work at low MICs against the entire ESKAPEE group of bacteria.

Peptide toxicity is a major hurdle in the development of therapeutic peptides. Hemolytic activity against red blood cells is one of the key factors to consider when evaluating peptide toxicity. High hemolytic activity can lead to anemia and other blood disorders, making highly hemolytic peptides unsuitable for pharmaceutical use. However, discovering low hemolytic peptides is a labor-intensive and time-consuming process that involves testing on mammalian red blood cells. To address this, we proposed a framework to identify low hemolytic therapeutic peptides.

All the frameworks we developed as part of different studies have also been made available as web applications. The wet-lab researchers can use these tools to narrow the search space while discovering low hemolytic AMPs active against different pathogens.