

**PREDICTION OF INSECTICIDE RESISTANT GENES *IN ANOPHELES GAMBIAE*  
USING A SEMI-SUPERVISED MACHINE LEARNING APPROACH**

**OWOLABI, JESUJOBA MARY  
(19PBF02177)**

**NOVEMBER, 2021**

**PREDICTION OF INSECTICIDE RESISTANT GENES *IN ANOPHELES GAMBIAE*  
USING A SEMI-SUPERVISED MACHINE LEARNING APPROACH**

**By**

**OWOLABI, JESUJOBA MARY  
(19PBF02177)**

**B. Sc (Hons), Microbiology, Bowen University, Iwo**

**A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE  
STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE  
AWARD OF MASTER OF SCIENCE (M.Sc.) DEGREE IN BIOINFORMATICS IN  
THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE  
OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY.**

**NOVEMBER, 2021**

## ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfillment of the requirements for the award of the degree of Masters of Sciences in Bioinformatics in the **DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY, OTA, NIGERIA.**

**Mr. John A. Philip**  
(Secretary, School of Postgraduate Studies)

-----  
*Signature and Date*

**Prof. Akan B. Williams**  
(Dean, School of Postgraduate Studies)

-----  
*Signature and Date*

## DECLARATION

I, **OWOLABI, JESUJOBA MARY (19PBF02177)** hereby declares that this research was carried out by me under the supervision of Prof. Ezekiel Adebisi of the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria. I attest that the dissertation has not been presented either wholly or partially for the award of any degree elsewhere. All sources of data and scholarly information used in this dissertation are duly acknowledged.

**OWOLABI, JESUJOBA MARY**

 4/27/2022

-----  
*Signature and Date*

## CERTIFICATION

We certify that this dissertation titled “**PREDICTION OF INSECTICIDE RESISTANT GENES IN *ANOPHELES GAMBIAE* USING A SEMI-SUPERVISED MACHINE LEARNING APPROACH**” is an original work carried out by **OWOLABI, JESUJOBA MARY (19PBF02177)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria under the supervision of Prof. Ezekiel Adebiyi. We have examined and found this work acceptable as part of the requirements for the award of Master of Science in Bioinformatics.

**Prof. Ezekiel F. Adebiyi**  
(Supervisor)

-----  
*Signature and Date*

**Prof. Olufunke O. Oladipupo**  
(Head of Department)

-----  
*Signature and Date*

**Prof. Folorunso Olusegun**  
(External Examiner)

-----  
*Signature and Date*

**Prof. Akan B. Williams**  
(Dean, School of Postgraduate Studies)

-----  
*Signature and Date*

## **DEDICATION**

To the Author and Finisher of My faith, who gave me the inspiration, strength and courage towards the achievement of this work. To my parents, spiritual father and siblings for their immense contribution and significant support towards the successful completion of this work.

## **ACKNOWLEDGMENTS**

I am indeed grateful to God for everything because without Him, my success is not guaranteed. My profound gratitude goes to my supervisor, Prof. Adebisi, co-supervisor, Dr. Titilope Dokunmu for all their tireless support during the course of this study.

I also want to acknowledge the HOD, Prof. Olufunke Oladipupo for her valuable counsel and contributions to this project. The department postgraduate coordinator, Dr. (Mrs.) Oni, for her tireless effort and contributions to this project.

I would like to express my gratitude to Dr. Isewon for her constructive criticism, Dr. Yagoub Adam, Mr. Aromolaran and every faculty member for their consistent interest in completion and success of this work.

My sincere and heartfelt gratitude goes to my parents, Pastor and Mrs. Owolabi for grooming me into a sort after asset and showing unconditional love and support from my very existence until now. To my wonderful, absolutely amazing siblings, your prayers and support rocked my world. The administrative staff of CApIC-ACE, thank you for your work. All ACE students and my colleagues (Fesobi Olumuyiwa, Rachael Apata, David Enoma, David Oladejo and Titilayo), you're absolutely amazing, from the love, to the mental support and the accolades, nobody could do it better than you all and thank you for the gift of friendship.

## TABLE OF CONTENTS

<b>CONTENT</b>	<b>Page</b>
<b>COVER PAGE</b>	<b>i</b>
<b>TITLE PAGE</b>	<b>ii</b>
<b>ACCEPTANCE</b>	<b>iii</b>
<b>DECLARATION</b>	<b>iv</b>
<b>CERTIFICATION</b>	<b>v</b>
<b>DEDICATION</b>	<b>vi</b>
<b>ACKNOWLEDGEMENTS</b>	<b>vii</b>
<b>TABLE OF CONTENTS</b>	<b>viii</b>
<b>LIST OF FIGURES</b>	<b>xii</b>
<b>LIST OF TABLES</b>	<b>xiv</b>
<b>LIST OF ALGORITHMS</b>	<b>xv</b>
<b>LIST OF ABBREVIATIONS AND ACRONYMS</b>	<b>xvi</b>
<b>ABSTRACT</b>	<b>xvii</b>
<b>CHAPTER ONE: INTRODUCTION</b>	<b>1</b>
1.1. Background information	1
1.2. Statement of the problem	3
1.3. Research questions	4
1.4. Aim and objectives of the study	4
1.5. Research methodology	4
1.6. Significance of the study	5
1.7. Scope of the study	5
1.8. Limitations of the study	6
1.9. Organization of the dissertation	6
<b>CHAPTER TWO: LITERATURE REVIEW</b>	<b>7</b>
2.1. Introduction	7
2.2. Overview of <i>Anopheles gambiae</i> resistance	7
2.3. Overview of machine learning models	8
2.4. Principles behind some supervised learning algorithm	10



2.4.1.	Decision Tree	10
2.4.2.	K-nearest neighbor	11
2.4.3.	Random Forest	12
2.4.4.	Naïve Bayes	12
2.4.5.	Support vector machine	13
2.5.	Probabilistic classifiers	14
2.5.1.	Calibration of a given model can be evaluated	15
2.5.2.	Methods for training a probabilistic classifier	16
2.5.3.	Examples of probabilistic classifiers	18
2.5.4.	Applications of different probabilistic classifiers	21
2.6.	Feature selection	21
2.7.	Overview of semi-supervised learning	23
2.7.1.	Semi-supervised classification method	24
2.7.2.	Graph-based methods	24
2.7.3.	Co-training	26
2.7.4.	Self-training	28
2.8.	Related works	30
<b>CHAPTER THREE: RESEARCH METHODOLOGY</b>		<b>33</b>
3.1.	Introduction	33
3.1.1.	Tools and Methods	34
3.1.2.	Python	34
3.1.3.	Sklearn	34
3.1.4.	Pandas	34
3.1.5.	Matplotlib	35
3.1.6.	Numpy	35
3.2.	Description of repositories	35
3.2.1.	Ensembl metazoa database	35
3.2.2.	g: Profiler	35
3.2.3.	Vectorbase	35
3.2.4.	Genecodis	36
3.2.5.	Deeploc	36
3.3.	Defining the gold standard	36

3.3.1.	Gold standard for insecticide resistance prediction	36
3.3.2.	Feature generation	37
3.3.3.	Amino acid composition	37
3.3.4.	Dipeptide composition	37
3.3.5.	Pseudo amino acid composition	37
3.3.6.	CTD	38
3.3.7.	DNA Sequence	38
3.3.8.	Subcellular localization	38
3.4.	Data preprocessing	39
3.4.1.	Feature selection	39
3.5.	Building the models	39
3.5.1.	Probabilistic classifiers	39
3.5.2.	Hyperparameter tuning	40
3.6.	Development of semi-supervised machine learning	42
3.7.	Performance evaluation of the models	43
3.8.	Performance measure of the model	44
3.9.	Biological Interpretation	45
<b>CHAPTER FOUR: RESULTS AND DISCUSSION</b>		<b>46</b>
4.1.	Introduction	46
4.1.1.	Dataset creation	46
4.1.2.	Result of Gold standard for resistance prediction	47
4.1.3.	Result of Data exploration and preprocessing	48
4.1.	Discussion of results for objective one	49
4.1.4.	Result of Feature selection	51
4.2.	Result of building and evaluation of the classifiers	52
4.2.1	Result of Performance Evaluation	54
4.3.	Result of self-training for insecticide resistance prediction	56
4.3.1.	Result of Comparative analysis of iterations between models	58
4.3.2.	Result of consensus-based voting	62
4.3.3.	Feature Importance of the models in python	63
4.2.	Discussion of results for objective two	65

4.3.4.	Result of functional enrichment analysis	67
4.3.	Discussion of results for objective three	90
4.3.1.	Discussion of functional enrichment analysis result of predicted genes	90
<b>CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS</b>		<b>93</b>
5.1.	Summary	93
5.2.	Conclusion	93
5.3.	Contribution to knowledge	94
5.4.	Recommendation	95
<b>REFERENCES</b>		<b>96</b>
<b>APPENDIX</b>		<b>114</b>
<b>Appendix A</b>		<b>114</b>
<b>Appendix B</b>		<b>132</b>

## LIST OF FIGURES

<b>Figure</b>	<b>Title of Figures</b>	<b>Page</b>
2.1	Overview of the type of machine learning	8
2.2	Support vector machine using decision boundary line	13
2.3	Feature selection methods	22
2.4	Graph-based semi-supervised learning	24
2.5	Co-training a multi-view learning method	27
3.1	Project activity workflow	33
3.2	Description of confusion matrix	45
4.1	Structure of the created dataset	46
4.2	Overall performance of machine learning models	52
4.3	Margin of confidence on Gradient Boosting Machine	53
4.4	Margin of confidence on XGBoost	54
4.5	Confusion matrix of Decision Tree on PIDD	55
4.6	ROC-AUC curve of Decision Tree on PIDD	55
4.7	Bar chart of probabilistic classifiers on the train set	57
4.8	Bar chart of probabilistic classifiers on the test set	57
4.9	Random Forest at each iteration	58
4.10	Logistic Regression at each iteration	59
4.11	XGBoost at each iteration	60
4.12	Gradient Boost Machine at each iteration	61
4.13	Confusion matrix of four models on the test set	62
4.14	Consensus prediction on the four models	63
4.15	Feature Importance in Logistic Regression	64
4.16	Feature Importance in Random Forest	64
4.17	A network cluster of genes and their functional annotations	71
4.18	Bar Plot showing functional enrichment result of the genes	72
4.19	Network cluster of the biological process in predicted genes	73

4.20	Word Cloud of the biological process in insecticide resistant genes	74
4.21	Network cluster of the cellular component of the predicted genes	75
4.22	Word cloud showing the cellular component of the predicted genes	76
4.23	Bar Plot showing the cellular component of the predicted genes	77
4.24	Network cluster showing the molecular function of predicted genes	78
4.25	Bar Plot showing the molecular function of the predicted genes	79
4.26.	Word cloud showing the molecular function involved in insecticide resistance	80
4.27	KEGG pathway of experimentally validated genes associated with increased Insecticide resistance in <i>Anopheles gambiae</i>	82
4.28	Network cluster of experimentally validated genes and associated pathways	83
4.29	Word cloud of experimentally validated genes	84
4.30	Bar Plot showing the molecular function of experimentally validated genes	85
4.31	Network cluster showing the molecular function of experimental genes	86
4.32	Bar Plot showing the biological process of the experimentally validated genes	87
4.33	Network cluster of the biological process of the experimentally validated genes	88
4.34.	Word cloud showing the biological process of the experimentally validated genes	88
4.35	Class probability estimation of a probabilistic classifier	88
4.36	Nested network of the predicted resistant genes	91

## LIST OF TABLES

<b>Table</b>	<b>Title of Tables</b>	<b>Page</b>
2.1	Summary of gaps in related works	32
3.1	Hyperparameter optimization in probabilistic classifiers and description	40
3.2	Description of PIDD diabetic dataset	43
4.1	Genes that represent the class labels for gene ontology dataset	47
4.2	Ranked features for information gain and chi square test	51
4.3	Accuracy scores for the selected features in mutual information and chi Square test	52
4.4	Comparative performance of the classification algorithm	54
4.5	Functional enrichment analysis of <i>Anopheles gambiae</i> predicted resistant genes (Top 20 molecular function)	68
4.6	Top 5 molecular function of predicted resistant genes and their orthologues in <i>Drosophila</i>	69
4.7	Functional enrichment analysis of <i>Anopheles gambiae</i> predicted resistant genes (Top 10 biological process)	70
4.8	Top 10 molecular function of experimentally validated genes of <i>Anopheles gambiae</i>	81

## LIST OF ALGORITHMS

Algorithm	Title of Algorithms	Page
Algorithm 1:	Pseudocode of Decision Tree	11
Algorithm 2:	Pseudocode of KNN	11
Algorithm 3:	Peudocode of Naïve Bayes	12
Algorithm 4:	Co-training pseudo-code	26
Algorithm 5:	Pseudo-code for a Classic Self-training	28
Algorithm 6:	Consensus-Voting code	42

## LIST OF ABBREVIATIONS AND ACRONYMS

AAC	Amino Acid Composition
ACHE	Acetylcholinesterase
CTD	Composition-transition-distribution
CPR	Cytochrome P450 reductase
CYP	Cytochrome
DPC	Dipeptide Composition
DT	Decision Tree
FN	False Negative
FP	False Positive
GABA	$\gamma$ -amino butyric acid
GBM	Gradient Boosting Machine
GO	Gene Ontology
GST	Glutathione S-transferase
HSP	Heat Shock Proteins
KDR	Knock-down resistant
KNN	K-nearest- neighbor
LR	Logistic Regression
MLP	Multilayer perceptron
NB	Naïve Bayes
RF	Random Forest
ROC-AUC	Area Under Receiver Operating Characteristic Curve
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VGSC	Voltage gated sodium channel
WHO	World Health Organization
XGBOOST	Extreme Gradient Boost



## ABSTRACT

Insecticides are used to manage insects that harm crops, livestock, and humans, as well as to eliminate pests that spread harmful infectious diseases. However, widespread use of insecticides, especially pesticides, has resulted in the reappearance of pest species that are totally resistant to more than two types of prescribed insecticides, resulting in an increase in global mortality rates. Insecticide resistance is defined as a heritable alteration in a pest population's sensitivity to insecticides, as evidenced by a repeated failure to achieve the expected degree of control when applied according to the level of recommended dosage. Experimental approaches have been extensively utilized in identifying resistance among many malaria vectors including *Anopheles gambiae*. However, these techniques used such as expression profiling and transcriptome analyses tends to be species specific, costly and time-consuming. Thus, computational technique for discovering resistant genes that is independent of species and cost-effective would aid in the advancement of insecticide resistance gene research. To this end, this research aims to identify other potential resistant genes with a self-trained semi-supervised approach using five probabilistic machine learning models such as Random Forest, Decision Tree, XGBoost, Gradient Boosting Machine and Logistic Regression. A total of 63 insecticide resistant genes were predicted across five models based on a consensus based voting scheme. With highly significant predictions, new insecticides can be formulated to counteract the activities of this predicted genes as functional analysis has shown their relationship to the already identified experimentally validated genes.

**Keywords:** *Anopheles gambiae*, Insecticide resistance, self-training, semi-supervised learning, biological interpretation