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 Adebiyi 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. Adebiyi, 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 counseland 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

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

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	γ-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