TRANSCRIPTIONAL SIGNATURE FOR TUBERCULOSIS PREDICTION USING ENSEMBLE LEARNING TECHNIQUE

BY

OKEZIE, ADAUGO FIONA

(18PCG01762)

B.Sc Computer Science, University of Nigeria, Nsukka, Enugu State

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 COMPUTER SCIENCE IN THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY.

DECEMBER, 2020

ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfillment of the requirements for the award of Master of Science (M. Sc.) degree in Computer Science in the Department of Computer and Information Science, College of Science and Technology, Covenant University, Ota, Ogun State, 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, OKEZIE, ADAUGO FIONA with matriculation number **18PCG01762,** herby declare that this dissertation **TRANSCRIPTIONAL SIGNATURE FOR TUBERCULOSIS PREDICTION USING ENSEMBLE LEARNING TECHNIQUE** was carried out by me under the supervision of Prof. Victor C. Osamor. This project is an original study in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria. All scholarly information used in this study is fully acknowledged..

OKEZIE, ADAUGO FIONA ______________________________

Signature and Date

CERTIFICATION

This is to certify that the dissertation titled "**TRANSCRIPTIONAL SIGNATURE FOR TUBERCULOSIS PREDICTION USING ENSEMBLE LEARNING TECHNIQUE**" is an original research work carried out by **OKEZIE, ADAUGO FIONA** with matriculation number **18PCG01762** under the supervision of Prof Victor .C. Osamor in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria.

Prof Victor C. Osamor ______________________________

(Supervisor) Signature and Date

Dr. Olufunke O. Oladipupo ______________________________

(Head of Department) Signature and Date

Prof. Olumide B. Longe ______________________________

(External Examiner) Signature and Date

Prof. Akan B. Williams ______________________________

Dean, School of Postgraduate Studies Signature and Date

DEDICATION

This project is dedicated to God Almighty for the grace and strength to complete the program and to my family for their love and support throughout my program.

ACKNOWLEDGEMENTS

I want to thank the almighty God for his sufficient grace to start this program and complete it successfully in due time. I appreciate him for the wisdom, knowledge, and understanding he gave me throughout the program.

I acknowledge and appreciate my parents Engr and Mrs. Ifeanyi Okezie, for their continuous support and love for me throughout my program. Their motivation and encouragement were a steady source of strength during the period. I also acknowledge the support I received from my siblings in all aspects.

Most importantly, to my supervisor Prof. V.C Osamor who has been an ideal thesis and project supervisor for me. His soothing advice, insightful criticism, and patient encouragement aided the writing of this thesis in innumerable ways. He has been more than just a supervisor to me throughout my attachment to him. I want to appreciate him and acknowledge his contributions.

In the department of Computer and Information Science, my gratitude and appreciation go to my HOD, Dr. Olufunke O. Oladipupo and to Prof. A.A. Azeta, for being patient and supportive throughout my entire program. My thanks also go to the entire Management of Covenant University for seeing that the vision has been fulfilled towards accomplishing great excellence and a thorough academic program.

I am grateful to the PG coordinator, Dr. Aderonke Oni, and all the faculty members of the CIS department for their productive remarks during the numerous presentations we had during the cause of this work. I acknowledge Dr. I. Odunayo for his significant contribution, support, and guidance during the project work and throughout my program. I also appreciate Dr. O. Emebo for been supportive throughout the course of my study. I am grateful.

Finally, my appreciation goes to my wonderful friends and collegues, who were with me throughout the program and encouraged me in one way or another. I appreciate them for all efforts made to ensure the program was stress-free.

I appreciate you all, and God bless you.

TABLE OF CONTENT

ACCEPTANCE **iii**

DECLARATION **iiv**

CERTIFICATION **v**

DEDICATION **vi**

ACKNOWLEDGEMENT **vii**

TABLE OF CONTENTS **viii**

LIST OF FIGURES **xi**

LIST OF TABLES **xii**

ABSTRACT **xiii**

CHAPTER ONE: INTRODUCTION

1.1 Background Information 1

1.2 Statement of The Problem 4

1.3 Aim and Objectives 5

1.4 Research Methodology 5

1.5 Justification of the Study 6

1.6 Organization of Dissertation 6

CHAPTER TWO: LITERATURE REVIEW

2.1 Tuberculosis 8

2.1.1 Life Cycle of Mycobacterium Tuberculosis 11

2.1.2 Types of Tuberculosis 12

2.1.3 Symptoms of Tuberculosis 13

2.1.4 Methods of Tuberculosis Diagnosis 13

2.1.5 Treatment of Tuberculosis 14

2.2 Bioinformatics and Omics 15

2.2.1 Bioinformatics 15

2.2.2 Omics 16

2.3 Biomarkers and Transcriptional Signature 18

2.3.1 Biomarkers 18

2.3.2 Uses of Biomarkers 21

2.3.3 Classification of Biological Markers 21

2.3.4 Biomarker Discovery 22

2.3.5 Steps in Biomarker Discovery 23
2.3.6 Biomarker Application Areas 24
2.3.7 Transcriptional Signature 27
2.4 Gene Expression Analysis 28
2.4.1 Transcription 28
2.4.2 Translation 29
2.5 Machine Learning 30
2.5.1 Supervised Learning 30
2.5.2 Unsupervised Learning 32
2.6 Ensemble Learning 33
2.7 Feature Extraction and Pre-Processing 35
2.7.1 Dimensionality Reduction 35
2.7.2 Feature Selection 36
2.8 Evaluation Metrics 38
2.8.1 Model Evaluation Techniques 38
2.8.2 Model Evaluation Metrics 39
2.9 Review Of Related Work 40
CHAPTER THREE: RESEARCH METHODOLOGY
3.1 Introduction 47
3.2 Methodology 48
3.2.1 Model Architecture 50
3.2.2 Dataset 51
3.2.3 Preprocessing and Normalization 52
3.2.4 Feature Selection 54
3.2.5 Model Building and Evaluation 55
3.2.5.1 Existing Weighted Voting Algorithm 57
3.2.5.2 Improved Weighted Voting Algorithm 58
CHAPTER FOUR: RESULT AND DISCUSSION
4.1 Introduction 61
4.2 The Implementation Tools Used 61
4.2.1 Machine Learning Model Development 61
4.3 System Requirements 63
4.3.1 Hardware Requirement 63
4.3.2 Software Requirement 63
4.4 Analysis of Results From Tuberculosis Expression Data Preprocessing And Feature Selection 65
4.4.1 Dimensionality Reduction Using PCA 65
4.4.2 Feature Selection Using RFE-CV 65

4.5 Model Evaluation 67

4.6 Overview 72

4.6.1 Transcriptional Signatures 73

4.6.2 Metrics and Accuracies of Individual Model 73

4.6.3 Metrics and Accuracies of Ensemble Technique 74

4.7 Scope and Limitations 75

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Summary of Findings 76

5.2 Conclusion 76

5.3 Contribution to Knowledge 77

5.4 Recommendations 77

REFERENCES 79

APPENDIX 87

LIST OF FIGURES

Figures	Title of Figures	Pages
2.1:	Shape of tuberculosis causing bacteria, bacilli	8
2.2:	Incidence rates of TB, 2018	9
2.3:	Estimated incidence of TB in 2018	10
2.4:	Tuberculosis in other parts of the body	10
2.5:	The Life cycle of Mycobacterium tuberculosis	11
2.6:	Progression of tuberculosis to various types	12
2.7:	An Ensemble Architecture	33
3.1:	Methodology Workflow	47
3.2:	Existing Model Architecture	49
3.3:	Model Architecture	51
4.1:	Rstudio IDE Interface	62
4.2:	Top 20 genes from PCA method	65
4.3:	Result of RFE-CV Feature selection	66
4.4:	Plot of RFE-CV result	67
4.5:	Visual Representation of KNN Classifier’s Confusion Matrix	68
4.6:	Visual Representation of SVM Classifier’s Confusion Matrix	69
4.7:	Visual Representation of NB Classifier’s Confusion Matrix	70
4.8:	Visual Representation of Existing Weighted Voting Ensemble Method Confusion Matrix	71
4.9:	Visual Representation of Existing Weighted Voting Ensemble Method Confusion Matrix	72

LIST OF TABLES

Table	Title of Table	Pages
3.1:	Phases and Output of Methodology	48
3.2:	Confusion Matrix	59
4.1:	Breakdown of Data Samples According to Classes	64
4.2:	Performance metrics used in the evaluation of the model	67
4.3:	Confusion Matrix Report for KNN	68
4.4:	Confusion Matrix Report for SVM	69
4.5:	Confusion Matrix Report for NB	70
4.6:	Confusion Matrix Report for Existing Weighted Voting Ensemble Method	71
4.7:	Confusion Matrix Report for Improved Weighted Voting Ensemble Method	72

ABSTRACT

Tuberculosis has the most considerable death rate among diseases caused by a single micro-organism type. The disease is a significant issue for most third world countries due to poor diagnosis and treatment potentials. Early diagnosis of tuberculosis is the most effective way of managing the disease in patients and reducing the mortality rate caused. Despite the several methods that exist in diagnosing tuberculosis, the limitations ranging from the cost in carrying out the test to the time taken to obtain the results have hindered early diagnosis of the disease. There is, therefore, the need to research alternative diagnostic methods that can aid diagnosis. Better testing with non-sputum samples, like blood, is now desirable and sustainable for diagnoses. Currently, blood transcriptional signatures (genes) are being considered since blood is easily accessible and can tell the state of the body at any point in time, and results can be gotten promptly. Lots of research will discover relevant transcriptional signatures to aid tuberculosis detection. These signatures can be easily observed from analyzing blood samples to know genes triggered in the body after tuberculosis-causing bacteria have infected it. This project work aims to develop a predictive model that would help in the diagnosis of TB and also identify relevant signatures that are affiliated with tuberculosis. The method used to carry out this research involved analyzing tuberculosis gene expression data obtained from GEO (Transcript Expression Omnibus) database and identifying relevant genes used to develop a classification model to aid tuberculosis diagnosis. A classifier combination of K.Nearest Neighbor, Bayes, and Support Vector Machine was used to develop the classification model. The weighted voting ensemble technique was used to improve the classification model's performance. The transcriptional signature obtained from the research includes "C4orf41", "GNPAT", "DHX15", "AGGF1", "ANKRD17", "TM2D1", "VAMP4". while the performance accuracy of the ensemble classifier was 0.95 which showed a better performance than the single classifiers which had 0.94, 0.92 and 0.87 obtained from KNN, SVM and NB respectively. The research clearly shows that the identified signatures can help in the early diagnosis of tuberculosis. The developed model can also assist health practitioners in the timely diagnosis of tuberculosis, which would reduce the mortality rate caused by the disease, especially in developing countries.

Keywords: Ensemble Learning, Weighted Voting Method, Tuberculosis, Machine learning Diagnosis, Predictive Model, Biomarkers