

**MAP REDUCE SECURITY MODEL FOR ASTHMA PREDICTION
IN CHILDREN USING FEDERATED XGBOOST**

**EKPO, RAPHAEL HENSHAW
(CUGP100277)**

APRIL, 2024

**MAP REDUCE SECURITY MODEL FOR ASTHMA PREDICTION
IN CHILDREN USING FEDERATED XGBOOST**

BY

**EKPO, RAPHAEL HENSHAW
(CUGP100277)**

**B.Sc Computer Science, University of Uyo, Uyo
M.Sc Computer Science, Covenant University, Ota**

**A THESIS SUBMITTED TO THE SCHOOL OF POSTGRADUATE
STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE AWARD OF DOCTOR OF PHILOSOPHY (Ph.D) DEGREE
IN COMPUTER SCIENCE, DEPARTMENT OF COMPUTER AND
INFORMATION SCIENCES, COLLEGE OF SCIENCE AND
TECHNOLOGY, COVENANT UNIVERSITY, OTA, OGUN STATE,
NIGERIA**

APRIL, 2024

ACCEPTANCE

This is to attest that this thesis is accepted in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria.

Miss Adefunke F. Oyinloye
(Secretary, School of Postgraduate Studies)

Signature and Date

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

Signature and Date

DECLARATION

I, **EKPO, RAPHAEL HENSHAW (CUGP100277)**, declare that this research work was carried out by me under the supervision of Prof. Victor C. Osamor of the Department of Computer Science and Prof. Ambrose A. Azeta of the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. I attest that this thesis has not been presented either wholly or partly for the award of any degree elsewhere. All the sources of data and scholarly information used in this thesis are duly acknowledged.

EKPO, RAPHAEL HENSHAW

Signature and Date

CERTIFICATION

This is to certify that the research work titled “**MAP REDUCE SECURITY MODEL FOR ASTHMA PREDICTION IN CHILDREN USING FEDERATED XGBOOST**” is an original research work carried out by **EKPO, RAPHAEL HENSHAW (CUGP100277)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria, under the supervision of Prof. Victor C. Osamor and Prof. Ambrose A. Azeta. We have examined and found this work acceptable as part of the requirements for the award of a degree of Doctor of Philosophy in Computer Science.

Prof. Victor C. Osamor
(Supervisor)

Signature and Date

Prof. Ambrose A. Azeta
(Co-Supervisor)

Signature and Date

Prof. Olufunke O. Oladipupo
(Head of Department)

Signature and Date

Prof. Adebukola S. Onashoga
(External Examiner)

Signature and Date

Prof. Akan B. Williams
(Dean, School of Post-Graduate Studies)

Signature and Date

DEDICATION

This thesis is dedicated to God almighty for His strength, and help throughout this study up to its successful completion.

ACKNOWLEDGEMENTS

I want to acknowledge the Almighty God, my strength, my protector, my helper, my provider and way maker who made this Ph.D a reality. Despite all the challenges, He gave me victory at last. To Him be all the glory, honour and praises.

I sincerely appreciate the presiding Bishop of Living Faith Church Worldwide and Chancellor of Covenant University, Dr. David Oyedepo. You are a blessing to our generation. I also express my profound gratitude to the Vice-Chancellor, Prof. Abiodun H. Adebayo, the Registrar, Mrs Regina A. Tobi-David, the Dean, School of Postgraduate Studies, Prof. Akan B. Williams and Dean, College of Science and Technology, Prof. Timothy Anake for their leadership role and support in this great citadel of learning.

I also wish to appreciate the effort of my supervisors, Prof. Victor C. Osamor and Prof. Ambrose A. Azeta, for their guidance, patience, and support throughout this PhD programme. I appreciate the head of the department, Prof. Olufunke O. Oladipupo. You have been a mother. My gratitude goes to the Departmental Post Graduate Coordinator, Dr Mrs. Itunu Isewon for her constant support. My special regards to my colleagues and all members of staff of the college of Natural and Applied Sciences, Crawford University. Thank you for standing by me.

I like to acknowledge, the Head of Department of Computer and Mathematical Sciences, Crawford University, Dr. Tayo Adefokun, for his understanding and encouragement. My heartfelt gratitude goes to my wife, Temitope, Raphael Henshaw, who has always been there for me, supported and encouraged me throughout the research work. My children Abundance Henshaw, Enoch Henshaw, and Asher Henshaw, who endured my absence for many days during this research work, you are the best. I am thankful to my brothers and sister in Christ – Dare Runsewe, Yusuf Kolawale and Yemisi Olabisi – for their constant prayers and support towards the success of this programme. I cannot forget the likes of Blessing Odede, Dr Adewole Adewummi, amongst others. These are the people who sacrificed their time to make this program possible. Finally, I am extremely grateful to my spiritual fathers, Pastor Uchechukwu Osinobi, and Pastor Matthew Ashimolowo for your prayers and teachings that were helpful to me in completing this programme.

TABLE OF CONTENTS

CONTENTS	PAGES
COVER PAGE	ii
ACCEPTANCE	iii
DECLARATION	iv
CERTIFICATION	v
DEDICATION	vi
ACKNOWLEDGEMENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xvi
ABSTRACT	xviii
CHAPTER ONE	1
INTRODUCTION	1
1.1 Background to the Study	1
1.2 Statement of the Problem	6
1.3 Aim and Objectives	7
1.4 Justification of the Research	8
1.5 Significance of the Study	9
1.6 The Scope of the Study	10
1.7 Glossary and Definition of Terms	10
1.8 Thesis Organization	11
CHAPTER TWO	12
LITERATURE REVIEW	12
2.1 Biology of Asthma	12
2.2 Asthma Predictive Factors	12
2.3 Asthma Adherence Monitoring Methods	13
2.4 Existing Asthma Predictive Models	16
2.5 Machine Learning	17
2.5.1. Machine learning Algorithm classification	20
2.5.2 Machine Learning Algorithms and Tasks	26
2.6 Imbalanced Classification	33
2.6.1 Methods for Handling Imbalanced Data Classification	35

2.7	Asthma Prognostic Factors from Published Works	37
2.8	Federated Learning	38
2.8.1	Taxonomy of Federated Learning	43
2.8.2	Categories of Federated Learning	44
2.8.3	Open Source Frameworks for Federated Learning	47
2.8.4	Federated Learning System Reference Architecture	60
2.8.5	Federated Learning Security Attack	62
2.8.6	Federated Learning Defence Strategy	68
2.8.7	Federated Learning Open Problems	74
2.9	Map Reduce Programming	85
2.9.1	Application Area of Map Reduce	87
2.9.2	Summary of Findings from the Literature Review	89
2.10	Symmetrical Uncertainty	89
2.11	Machine learning classification approach for asthma prediction models in Children	91
2.11.1	Search Strategy	93
2.11.2	Selection Process	94
	CHAPTER THREE	114
	METHODOLOGY	114
3.1	Introduction	114
3.2	Overview of the Proposed Model	117
3.3	Components of the Proposed Model	118
3.3.1	Component of the Federated Model	119
3.3.2	Problem Definition	121
3.3.3	Model Formulation	123
3.3.4	Security of the Proposed Model	124
3.3.5	Enhancing the Traditional Map Reduce Model for Parallel Preprocessing and Data Locality	127
3.3.6	Problem Definition of the Symmetrical Uncertainty and Normalisation Interaction Gain	128
3.4	Experimental Setup	130
3.5	Dataset Source	131
3.5.1	Description of the Nigerian Hospital Record System (HRS)	131
3.5.2	Description of the National Survey of Children Health (NSCH) Dataset	134
3.6.	Data Duplication of the Nigerian Hospital Record System (HRS)	134

3.6.1	Dropping of Irrelevant samples from National Survey of Children Health Dataset	135
3.6.2	Choice of Programming Language	136
3.7.	Evaluation Overview	137
	CHAPTER FOUR	140
	RESULTS AND DISCUSSION	140
4.1	Introduction	140
4.2	Exploratory Data Analysis of the Nigerian Hospital Dataset	140
4.3.	Data Preprocessing of the Nigerian Hospital Record System	143
4.4.	Dealing with Missing Values of the Nigerian Dataset	144
4.5	Dropping of Irrelevant Samples from Nigerian Hospital Dataset	146
4.5.1.	Exploratory Data Analysis of the National Survey of Children Health (NSCH) Dataset	147
4.5.2	Preprocessing and Symmetric Uncertainty of the NSCH Dataset	149
4.6	Results of the Implemented Model on the Nigerian Hospital Dataset	150
4.6.1	Classification Report of the Implemented Model on the Nigerian Hospital Dataset	151
4.6.2	Confusion Matrix of the Implemented Model on the Nigerian Hospital Dataset	151
4.6.3	Confusion Matrix of the Implemented Model on the Nigerian Hospital Dataset	152
4.6.4	Actual Versus Predicted Result of the Implemented Model on the Nigerian Hospital Dataset.	154
4.6.5	Actual Versus Predicted Result of the Implemented Model on the Nigerian Hospital Dataset	154
4.6.6	Results of the Implemented Model on the NSCH Dataset	155
4.6.7	Classification Report of the Implemented Model on the NSCH Dataset	155
4.7.	Confusion Matrix of the Implemented Model on the National Survey of Children Health (NSCH) Dataset	156
4.7.1.	ROC AUC of the Implemented Model on the National Survey of Children Health (NSCH) Dataset	157
4.7.2	Precision-Recall Curve of the Implemented Model on the National Survey of Children Health (NSCH) Dataset	158
4.7.3.	Actual Versus Predicted Result of the Implemented Model on the National Survey of Children Health (NSCH) Dataset	159
4.8.	Benchmarking of the Proposed Model with the Existing Model in Literature	160

4.9	Simulated Attack Result of the proposed Map Reduce Model and without Map Reduce	161
CHAPTER FIVE		164
SUMMARY, CONCLUSION AND RECOMMENDATIONS		164
5.1	Summary of Findings	164
5.2	Conclusion	165
5.3	Contribution to Knowledge	166
5.4	Recommendations	166
5.5.	Limitation of the Study	167
5.6.	Areas of Further Research	167
REFERENCES		168

LIST OF TABLES

TABLES	LIST OF TABLES	PAGES
2.1	Publish Asthma prognostic factors	38
2.2	Taxonomy of Security 2Table 2.2 Taxonomy of Security Attacks in Learning	62
2.3	Features of TEE	72
2.4	Related Works	77
2.5	Search Strategy Result	93
2.6	Summary of Machine Learning Classification Approach Methodologies and Performance Metric Values for Asthma Prediction in Childre	102
3.1	Objective and Research Methodology Mapped Out	116
3.2	Dataset Description	132
3.3	Target Class of the National Survey of Children Health Dataset	134
3.4	Dataframe After Dropping Dropping Duplicate Observation	134
3.5	Dataset After Removing Disease that are not Related to the Study	136
4.1	Head Section of the Dataset of the Nigeria Hospital Dataset	141
4.2	Output of the Dataset Information	142
4.3	Result of the Describe Function	143
4.4	Output of the Target Class Filtering	144
4.5	Percentage of Missing Values 1	146
4.6	Data Head of the National Survey of Children Health Dataset	148
4.7	Dataset After Filtering Patients between the Ages of 0-6	149
4.8	DataFrame After Preprocessing and Symmetric Uncertainty	150
4.9	Result of the Nigeria Hospital Centralized Model	150
4.10	Classification Report of the Nigeria Hospital Centralized Model	151
4.11	Actual versus Predicted Result	154
4.12	Federated Averaging of the Nigeria Hospital Dataset	154
4.13	Result of the Centralized Model on the NSCH Dataset	155

4.14	Classification Report of the Implemented Model on the NSCH Dataset	156
4.15	Federated Averaging of the NSCH Dataset	159
4.16	Benchmarking of the Performance Accuracy of the Proposed Model with Existing Models in Literature	161
4.17	Proposed Model Result with Existing Models in the Nigeria Hospital Dataset	163
4.18	Comparing Against Map Reduce	163

LIST OF FIGURES

FIGURES	LIST OF FIGURES	PAGES
2.1	Types of Machine Learning : Source:(Sarker, 2021)	20
2.2	A Typical Structure of Machine Learning Based Prediction Model	21
2.3	Supervised Learning Model Structure	22
2.4	Unsupervised Learning Model Structure	25
2.5	Federated Learning Architecture Client-Server	39
2.6	Architecture of Federated AI Technology Enabler (FATE)	48
2.7	PySyft Architecture	50
2.8	TensorFlow Federated Architecture	52
2.9	Paddle Federated Learning (PFL) Architecture	54
2.10	Federated Learning and Differential Privacy (FL & DP) Architecture	55
2.11	Open Federated Learning (Open FL) High-Level Overview	58
2.12	Mapping of Security Attack Counter Measures	63
2.13	Diagram of Map Reduce	86
2.14	Differences between the conventional Approach & Map Reduce Technique	87
2.15	PRISMA flow Diagram for the Study	95
3.1	The Research Workflow	115
3.2	Centralize Model of the Proposed Asthma Predicted Model	118
3.3	Component of the Federated Model	120
3.4	Data Preprocessing Workflow	121
3.5	Flowchart of the Symmetrical Uncertainty Feature Selection	129

3.6	Google Colab on Cloud for Training Deep Learning Models	131
3.7	Number of Unique Target Class of the NCSH Dataset	135
3.8	Pie Chart Distribution of Negative and Asthma Label	136
3.9	Confusion Matrix	139
4.1	Missing Values of Features	145
4.2	Number of Unique Target Class	147
4.3	Confusion Matrix of the Implementation Model	152
4.4	AUCROC Score of the Implemented Model	153
4.5	Precision-Recall Curve of the Implemented Model	153
4.6	Confusion Matrix of the Implemented Model in the NSCH Dataset	157
4.7	AUC RUC Curve of the Implemented Model on the NSCH Dataset	158
4.8	Precision-Recall Curve of the Implemented Model on the NSCH Dataset	159
4.9	Benchmarking of the Performance Accuracy of the Proposed Model with Existing Models of Literature	161
4.10	Evaluated Privacy Analysis of the Implemented Model	162

LIST OF ABBREVIATIONS

ACC	Accuracy
ADB	Adaboost
ANN	Artificial Neural Network
AUC	Area Under the Receiver Operating Curve
BP	Back Propagation
BPANN	Back Propagation Neural Network
CAPE	Children Asthma Prediction in Early Life
CAPP	Children Asthma Prediction at Preschool
CPDS	Cluster Primal Dual Splitting
CPX	Cox Proportional Hazard
CSAMM	Context sensitive Auto Association Memory Neural Network Model
CVAE	Conditional Variational Auto Encoder
DNN	Deep Neural Network
DT	Decision Tree
HER	Electronic Health Record
FCB	Fast Correlation- Based Filter
FENO	Fractional Exhaled Nitric Oxide
FEV1	Forced Exploratory Volume in one seconds
GA	Genetic Algorithm
GBDT	Gradient Boosting Decision Tree
GBM	Gradient Boost Machine
GOOGLE COLAB	Google Collaboratory
HRS	Hospital Record System
HSCRIP	High Sensitivity Creative Protein
IDE	Integrated Development Environment
KNN	K- Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Logistic Regression
LOAC	Level of Asthma Control
LR	Linear Regression
LR	Logistic Regression
LSSVM	Least Square Support Vector Machine
LSTM	Long Short erm Memory

MCC:	Matthews Correlation Coefficient
MDBS:	Multi Boost with Decision Support
MEF50:	Maximum Exploratory Flow of 50% of the vital flow capacity
MLP:	Multi Layer Perceptron
MLPNN:	Multi Layer Perceptron Neural Network
MNL:	Multinomial Logistic Regression
NB:	Naïve Bayes
NSCH:	National Survey of Children Health
PR:	Poission Regression
PRE:	Precision
PSO:	Particle Swarm Optimization
RBF:	Radial Basis Function
RCC:	Recall
RDSVM:	Radial Bias kernel function Support Vector Machine
RF:	Random Forest
RFE:	Recursive Feature Elimination
SEN:	Sensitivity
SME:	Security of Multiparty Communication
SPE:	Specificity
SSVM:	Sparse Support Vector Machine
SVM:	Support Vector Machine
TEE:	Trusted Executive Environment
XGB:	Extreme Gradient boosting

ABSTRACT

A substantial number of asthma development prediction models in children, such as conventional methods involving risk factors, logistic regression, the hybrid of statistical methods, and machine learning based approaches, exist. The problem associated with conventional methods of asthma prediction in children is low predictive accuracy of the model. However, using centralised machine learning approaches in healthcare requires training the learning models on large datasets. Besides cost, data privacy and security represent the main problems with centralised machine learning. The objective of this study is to develop a Map Reduce Security for asthma prediction in children using a federated XGBoost as a response to the aforementioned limitations associated with the existing asthma prediction model for children. This study leveraged two diverse datasets: the Nigerian Hospital Asthma dataset and the National Survey of Children's Health dataset for benchmarking purposes. After preprocessing the dataset, the symmetrical uncertainty and normalization interaction gain, as well as the undersampling approach, were employed for feature selection and dataset balancing. The system was trained, and tested using Federated Artificial Intelligence (AI) Technology Enabler (FATE) and extended to XGBoost model with one central server for federated algorithm averaging. The map reduce security measure was employed for the input data during training to avoid data leakage. The study was implemented with the Python programming language on Google Collaboratory (Collab) environment. The results of the analysis showed considerable high accuracy of 0.98, precision (0.98%), recall (0.98%), and F1-score (0.99%) for the asthmatic class and precision of 0.98%, as well as F1-score of 0.99% for the non-asthmatic class. The implemented model was benchmarked with the existing study on asthma prediction model in children using the same NSCH dataset of 23 features and 50212 samples. This study achieved a prediction accuracy of 93.8%, outperforming the existing models. The simulated attack on the implemented model results show that the model correctly identified whether a data point was used in the training process with predictive accuracy of 97.94%, membership inference attack, and data loss of 0.014%, respectively. In conclusion, the study finds that the model developed by the researchers performed better in predicting childhood asthma than some of the state-of-the-art machine learning algorithms.

Keywords: Asthma prediction model, Asthma in Children, Encryption, Federated Learning, Performance Metrics and Machine Learning.