

**Development of an Extended Biomedical Named Entity Recognition and  
Relation Extraction Model for Malaria using BioBERT**

**AJAO, ADEMOLU DANIEL  
(22PCG02416)**

**BSc Computer Science, Mountain Top University, Ogun State**

**AUGUST, 2024**

**Development of an Extended Biomedical Named Entity Recognition and  
Relation Extraction Model for Malaria using BioBERT**

**BY**

**AJAO, ADEMOLU DANIEL  
(22PCG02416)**

**BSc Computer Science, Mountain Top University, Ogun State.**

**A DISSERTATION SUBMITTED TO THE SCHOOL OF  
POSTGRADUATE STUDIES IN PARTIAL FULFILMENT OF THE  
REQUIREMENT OF THE MSc DEGREE IN COMPUTER SCIENCE  
(SOFTWARE ENGINEERING) OF THE DEPARTMENT OF  
COMPUTER AND INFORMATION SCIENCES, COLLEGE OF  
SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY, OTA,  
OGUN STATE, NIGERIA**

**AUGUST, 2024**

## **ACCEPTANCE**

This is to attest that this dissertation is accepted in partial fulfilment of the requirements for the award of the degree of Master of Sciences in Computer Science in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, 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 declare that **AJAO, ADEMOLU DANIEL (22PCG02416)**, carried out this research entitled **“DEVELOPMENT OF AN EXTENDED BIOMEDICAL NAMED ENTITY RECOGNITION AND RELATION EXTRACTION MODEL FOR MALARIA USING BIOBERT”**. It was carried out under the supervision of Dr. Oluranti Jonathan, Concepts of this research project are results of the research carried out by Ajao, Ademolu Daniel and ideas of other researchers have been fully recognized.

**AJAO, ADEMOLU DANIEL**

**Signature and Date**

## **CERTIFICATION**

This is to certify that this dissertation titled “**DEVELOPMENT OF AN EXTENDED BIOMEDICAL NAMED ENTITY RECOGNITION AND RELATION EXTRACTION MODEL FOR MALARIA USING BIOBERT**” is original research carried out by **AJAO, ADEMOLU DANIEL (22PCG02416)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota. Ogun State, Nigeria, under the supervision of Dr. Oluranti Jonathan. We have examined and found this work acceptable as part of the requirements for the award of Master of Science (M.Sc.) in Computer Science.

**Dr. Oluranti Jonathan**

**(Supervisor)**

**Signature and Date**

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

**Signature and Date**

**Prof. Zacchaeus O. Omogbadegun**  
**(External Examiner)**

**Signature and Date**

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

**Signature and Date**

## **DEDICATION**

I dedicate this project to God Almighty for His Help and ever-sufficient Grace, Wisdom and Knowledge given to me throughout my Master's Degree Programme. My Family and Friends for their words of encouragement and support.

## ACKNOWLEDGEMENTS

I use this medium to appreciate the following people/organizations that have supported me in one way or another during my study.

First and foremost, I would like to thank my project supervisor, Dr. Oluranti Jonathan, for his invaluable guidance, insightful feedback, and constant encouragement throughout the project.

Secondly, I want to thank the Head of the Department of Computer and Information Sciences, Prof. Olufunke Oladipupo, for giving me the unique privilege to pursue my interests and for her guidance.

Also, I would like to thank the faculty members for their feedback and advice and for giving expert opinions that helped shape me in the right direction.

I also want to thank CApiC-ACE for sponsoring, supporting and giving me the privilege to further my studies for this degree.

I also want to appreciate my parents, Mr Adefemi and Mrs Florence Ajao and my siblings for their support and encouragement that gives me the courage to carry on towards the journey of excellence

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGES</b>
<b>TITLE PAGE</b>	<b>i</b>
<b>COVER 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>xi</b>
<b>LIST OF TABLES</b>	<b>xiii</b>
<b>ABBREVIATIONS</b>	<b>xiv</b>
<b>ABSTRACT</b>	<b>xv</b>
<b>CHAPTER ONE: INTRODUCTION</b>	<b>1</b>
1.1. Background to the Study	1
1.2. Statement of the Problem	3
1.3. Aim and Objectives of the Study	4
1.4. Research Methodology	4
1.5. Significance of the Study	5
1.6. Scope of the Study	5
1.7. Organization of the Dissertation	5
<b>CHAPTER TWO: LITERATURE REVIEW</b>	<b>7</b>
2.1. Preamble	7
2.2. Large Language Models	7
2.2. Natural Language Processing	10
2.3. Biomedical Models and Large Language Models	11
2.4. Natural Language Processing in the Biomedical Domain	24
2.5. BERT and Named Entity Recognition	25
2.6. Relation Extraction	27
2.7. Applications of the Biomedical Language Model in Relevant Studies	29
2.8. Related Works	33



2.9.	Summary of Findings	34
<b>CHAPTER THREE: METHODOLOGY</b>		<b>35</b>
3.1.	Preamble	35
3.2.	Models Overview	35
3.2.1.	BioBERT	35
3.2.2.	Multi-BioNER	36
3.2.3.	FT-BioBERT	37
3.3.	Model Architecture	37
3.3.1.	Long Short-Term Memory (LSTM) Layer	37
3.3.2.	Random Forest	40
3.3.3.	Support Vector Machine (SVM)	40
3.3.4.	Gradient Boosting Machine (GBM)	42
3.4.	Data Collection	43
3.4.1	Sources of Data	44
3.4.2	Datasets for Model Analysis	44
3.5.	Data Pre-Processing	45
3.5.1.	Data Pre-Processing for our Model	45
3.5.2.	Initial Data Cleaning	47
3.5.3.	Tokenization and Annotation Alignment	48
3.5.4.	Handling Missing Data	48
3.6.	Data Splitting	48
3.6.1.	Training Set (70%)	49
3.6.2.	Validation Set (15%)	49
3.6.3.	Test Set (15%)	49
3.7.	Model Training and Evaluation	49
3.7.1.	Training Setup	49
3.7.2.	Performance Evaluation	50

<b>CHAPTER FOUR: RESULTS AND DISCUSSION</b>	<b>52</b>
4.1. Preamble	52
4.2. Results on BC5CDR-Disease Dataset	52
4.2.1. Results on BC5CDR-Disease Dataset for System 1:	52
4.2.2. Results on BC5CDR- Diseases for System 2	53
4.2.3. Analysis of Results	54
4.3. Results on BioRED dataset	55
4.3.1 Results on BioRED Dataset on System 1	55
4.3.2. Results on BioRED Dataset on System 2	56
4.3.3. Analysis of Results	56
4.4. Results on NCBI-Disease Dataset	57
4.4.1 Results on NCBI-Disease Dataset on System 1	57
4.4.2. Results on NCBI-Disease Dataset on System 2	58
4.4.3 Analysis of Results	59
4.5. Summary of Results	59
4.6. Discussion of Results	60
<b>CHAPTER FIVE: RECOMMENDATIONS AND CONCLUSION</b>	<b>62</b>
5.1. Summary of Findings	62
5.2. Contributions to Knowledge	63
5.3. Limitations of the Study	64
5.4. Recommendations for Future Research	64
5.5. Conclusion	65
<b>REFERENCES</b>	<b>67</b>

## LIST OF FIGURES

FIGURE	TITLE OF FIGURE	PAGE
Figure 1.1:	A basic flow diagram depicting various stages of LLMS from pre-training to prompting/utilization (source: Naveed <i>et al.</i> , 2024)	2
Figure 1.2:	Core Concepts of Natural Language Processing (source: Geetha <i>et al.</i> , 2023)	3
Figure 2.1:	A 4th generation compiler based on large language models (source: Marcondes <i>et al.</i> 2023)	8
Figure 2.2:	LLM's use in Healthcare (source: Reddy, 2023)	9
Figure 2.3:	The NLP pipeline for Smart Healthcare (source: Zhou <i>et al.</i> 2024)	10
Figure 2.4:	The overall Architecture of keBioLM (source: Yuan <i>et al.</i> 2021)	13
Figure 2.5:	Pre-training and Fine-tuning of BioALBERT on NER (Source: Naseem <i>et al.</i> ,2021)	14
Figure 2.6:	Overview of the proposed integration process and model architecture. (source: Arabzadeh and Bagheri, 2023)	15
Figure 2.7:	The ABioNER model (source: Boudjellal <i>et al.</i> ,2021)	16
Figure 2.8:	Application of BERT to perform BioNER in the MRC framework framework (Source: Sun <i>et al.</i> , 2021)	17
Figure 2.9:	Architecture of the BiLSTM Model (source: Naseem <i>et al.</i> , 2020)	18
Figure 2.10:	The KEBLM model (source: Lai <i>et al.</i> , 2023)	19
Figure 2.11:	Proposed Neural Model (source: Narayanan <i>et al.</i> , 2022)	21
Figure 2.12:	The Architecture of the BioBIT model (source: Buonocore <i>et al.</i> , 2023)	22
Figure 2.13:	StaResGRU-CNN structure (Source: Ni <i>et al.</i> , 2021)	23
Figure 2.14:	An Overview of BioMedBERT Information Retrieval Architecture (Source: Chakraborty <i>et al.</i> , 2020)	24
Figure 2.15:	Overall Pre-Training and Fine-Tuning Processes of BERT (Source: Lee <i>et al.</i> , 2019)	25
Figure 2.16:	An overview of BioVocabBERT Tokenizer (Source: Gutiérrez <i>et al.</i> , 2023)	26
Figure 2.17:	The Triplet linearization of REBEL (Source: Huguet Cabot and Navigli, 2021)	28

Figure 2.18: The architecture of the model with type-aware map memories (TaMM) (Source: Chen <i>et al.</i> ,2021)	29
Figure 2.19: Conditional Knowledge Infusion into Pretrained Language Models (Source: Jha and Zhang, 2022)	30
Figure 2.20: Querying Knowledge Bases (KB) and language models (LM) with EHR note context to enhance factual knowledge extraction. (Source: Yao <i>et al.</i> 2022)	32
Figure 3.1: Biobert Model (Source: Lee <i>et al.</i> 2019)	36
Figure 3.2: Multi-BioNER Model (Source: Park <i>et al.</i> , 2023)	37
Figure 3.3: Ft-BioBERT Architecture	43
Figure 3.4: The Search Query Result as of 14 <sup>th</sup> May 2024	44
Figure 3.5: Documents Retrieved after the first iteration of Preprocessing	46
Figure 3.6: Second round of Data Preprocessing	46
Figure 4.1: BC5CDR results for System 1	53
Figure 4.2: BC5CDR-Disease Dataset Results on System 2	54
Figure 4.3: BioRED Dataset Results on System 1	55
Figure 4.4: BioRED results on System 2	56
Figure 4.5: NCBI-Disease Results on System 1	58
Figure 4.6: NCBI-Disease on System 2	59

## LIST OF TABLES

<b>TABLE</b>	<b>TITLE OF TABLE</b>	<b>PAGE</b>
Table 4.1:	Result of BC5CDR-Disease Dataset On System 1	52
Table 4.2:	Result of BC5CDR-Disease Dataset On System 2	53
Table 4.3:	BioRED Dataset Results on System 1	55
Table 4.4:	BioEDResults on System 2	56
Table 4.5:	Results On NCBI-Disease Dataset On System 1	57
Table 4.6:	Results On NCBI-Disease Dataset On System 2	58

## ABBREVIATIONS

Named Entity Recognition	NER
Relation Extraction	RE
Large Language Model	LLM
Biomedical Bidirectional Encoding Representations from Transformers	BioBERT
Natural Language Processing	NLP
Convolutional Neural Network	CNN
Kernelized Support Vector Machine	KSVM
Biomedical Information Extraction	BioIE
Biomedical Natural Language Processing	BioNLP
Biomedical Named Entity Recognition	BioNER
Long Short-Term Memory	LSTM
Random Forest	RF
Support Vector Machine	SVM
Gradient Boosting Machines	GBM
Bidirectional Encoder Representations from Transformers	BERT

## ABSTRACT

This study aims to enhance the biomedical Named Entity Recognition and Relation Extraction model for use in the malaria subdomain by fine-tuning the existing BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) model. The process of fine-tuning involves adjusting the parameters of the BioBERT model to suit the characteristics of the malaria subdomain better. The model is intended to improve the process of recognizing entities and their relationships in the context of malaria-associated publications. This solves an essential problem in connection with the inapplicability of the previously developed models in the biomedical field. The study uses complex and highly effective machine learning algorithms, such as Long Short-Term Memory, Random Forests, Support Vector Machine and Gradient Boosting Machine, to fine-tune the existing BioBERT model, leading to the FT-BioBERT model. The fine-tuned model is compared with other models, such as BioBERT and Multi-BioNER, over three datasets, namely BC5CDR-Disease, BioRED, and NCBI-Disease. The fine-tuned model achieved notable performance improvements: achieving 92.4% in accuracy, which is a 3.13% increase from BioBERT and 2.33% from Multi-BioNER and attaining 91.8% in precision, 92.7% in recall, and 92.2% F1-score which is a of 3.15% improvement over BioBERT, and 2.23% improvement over Multi-BioNER. Based on the results, we confirm that the proposed model can effectively identify and extract entities and their relationships when supplied with malaria literature and, therefore, is suitable for biomedical text mining. We hope that the study's findings will provide new avenues that will lead to the creation of domain-related NLP applications in malaria-related fields.

***Keywords: Named Entity Recognition, Relation Extraction, BioBERT, Biomedical Language Model, Biomedical Natural Language Processing***