**TABLE OF CONTENTS**

**CHAPTER ONE**

**INTRODUCTION**

* 1. General Introduction ………………………….. 3
  2. Problem Definition …………………………….. 4
  3. Aim & Objectives of Study …………………... 5
  4. Research Justification …………………………. 5
  5. Scope and Limitation of Study ……………… 6
  6. Definition of Terms …………………………….. 7
  7. Chapter Layout ………………………………….. 7

C**HAPTER TWO**

**LITERATURE REVIEW**

* 1. Introduction
  2. Brief History of Study Area (where applicable)
  3. Review of relevant literature and state-of-the-art
  4. Proposed Contribution

C**HAPTER THREE**

**SYSTEM DESIGN METHODOLOGY**

* 1. Introduction
  2. System Requirements Specification
  3. System Design (OO or Structured – must be explained)
     + 1. Logical Design
       2. Input Design (where applicable)
       3. Output Design (where applicable)
       4. Use Case Diagram
       5. Activity Diagram
       6. Class Diagram
     1. Physical Design
        1. Program Specification
        2. Layout of Tables Design & Database Structure (where applicable)
        3. System Controls (e.g. security, input, output)

C**HAPTER FOUR**

**SYSTEM IMPLEMENTATION**

* 1. Introduction
  2. Features and Choice of Implementation Language
  3. System Testing Strategies
     1. Unit Test
     2. Integration Test
  4. Target Computer System Requirements
  5. Software Maintenance Issues
     1. Corrective Maintenance
     2. Preventive Maintenance
     3. Adaptive Maintenance

C**HAPTER FIVE**

**SUMMARY AND CONCLUSION**

* 1. Introduction
  2. Summary
  3. Recommendations (Future Research)
  4. Conclusion

REFERENCES

APPENDICES

Appendix A: (any attachments where applicable)

Appendix B: Source Code Listing

Appendix C: Sample Output

CHAPTER ONE

INTRODUCTION

1.1 GENERAL INTRODUCTION

Today, in Nigeria, medical diagnosis are carried out by doctors, specialized in their field of study. Through proper and thorough examination of patients and their organs, these doctors are able to identify symptoms of known diseases. Their findings, aids them in the diagnosis of a patient and further prescription of treatment for the diagnosis.

For most diseases, the symptoms are finite and consistent in all occurrence or instances. For example, Heart disease is characterized by chest pain along with other symptoms, therefore patients who complain of chest pain are likely to be diagnosed of heart disease.

“Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change, when exposed to new data.” (whatis.techtarget.com, 2016).

Machine learning is basically algorithms that can learn from observational data and can make predictions based on it. The learning algorithms are fed these data and a model is inferred by mathematically determining the relationship between the data and the predicted values. This model in turn is used to predict new values from new data supplied to it.

“A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves it’s performance P at task T, following experience E.” (Mitchell, 2006)

With the various advancements in machine learning algorithms, the task of learning from existing curated data has been made possible. This has given rise to various implementation of machine learning algorithms in different fields e.g. Medical Diagnosis, Natural language processing, Recommender systems e.t.c.

Machine learning algorithm are classified into three different categories based on their method of learning and output, they include the following:

1. Supervised Learning:

Supervised learning involves the provision of data attributes and already observed corresponding result of the data attributes as input or training data and the algorithm associates attributes to the result to form a model which will predict any other data attributes supplied to it.

2. Unsupervised Learning:

In unsupervised learning, the model is given raw observational data to make sense out of it. It groups the data into “clusters” or groups identified by similarity metric which might not conform to predisposed stereotypes. This type of learning are suitable in fields like Speech Recognition.

3. Reinforcement Learning:

Reinforcement Learning deals with dynamic data (data that changes). It tries to make decisions based on these data, it is provided with negative or positive feedback on every decision made. This feedback guides it’s subsequent decisions.

1.2 PROBLEM DEFINITION

Computers are made to ease and aid human daily activities, by automating tasks. These task varies in complexity, and since the invention of computers, great progress have been made in simplifying them. One of these tasks includes medical diagnosis. Diagnosis are carried out through thorough examination of patients by qualified medical staffs and Doctors, and their expert knowledge is used to prescribe possible solutions for the diagnosis.

This process of diagnosis i.e gathering data through examination, and using this data to proffer diagnosis, can be achieved by machine learning. There are various algorithms to choose from, to perform this diagnosis. Each one of the available algorithms have their strengths and limitations, which varies with their different implementations.

For a task as delicate and sensitive as medical diagnosis, there is little tolerance for errors and inefficiency, as the aim of using computers to aid human tasks is to increase efficiency and and minimize errors to its barest minimum.

The process of preparing a machine learning model for prediction, involves various steps which includes;

1. Collecting data.

2. Preparing the data.

3. Choosing the right algorithm.

4. Training the algorithm.

5. Testing the algorithm.

Of all the above mentioned steps, the first three are the most important, as they determine the outcome of the last two as well as the performance of the model. These steps are rigorous activities that takes a lot of time and difficulty to get right.

In Nigeria, due to little or no existing collection of data to carry out machine learning, the first three steps are made even more difficult. But with proper preparation of available data, an optimal machine learning model can be made for medical diagnosis.

1.3 AIM AND OBJECTIVES OF STUDY

The aim of this project was to go through the entire process involved in machine learning, to illustrate the thorough procedure in preparing available data and choosing the right algorithm. This process is a thorough process as there is little tolerance for errors.

Every machine learning algorithm and model is only as good as it’s data. Wrong data supplied to the algorithm for training, will definitely render it non-reliable. Data to be used for machine learning has to be normalized for effective usage by the algorithm. This process of normalization is carried out the best possible way, to limit errors and inconsistency.

The objective is to have a working model capable of predicting if a person has a heart disease or does not. Heart disease are of different types, e.g Rheumatic heart disease, Hypertensive heart disease, Ischemic heart disease etc. This project emphasizes on the presence of heart disease of any kind.

This whole process would not only come up with the best possible form, a data set should be in, for training a machine learning algorithm capable of predicting the presence of heart disease, it would also test some various available algorithms, to infer the best and suitable algorithm for predicting the presence of heart disease.

Some algorithms work well with problems that involves classified data, while others are suitable for continuous data and others for binary data. The problem of predicting the presence of heart disease is a binary data, i.e it is either present (true / 1) or absent (false / 0)

1.4 RESEARCH JUSTIFICATION

There are very few research on machine learning which show a comprehensive account of how data was prepared for a machine learning process. Most of these research productions, get prepared data, supply them to algorithms for training, and tests them.

The process of data preparation is a very important one. Data is the primary dictator of the outcome of a machine learning research or implementation. Without enough, correct data, the result of the machine learning process, is deemed biased and flawed.

What can be wrong with data? There is a hierarchy of problems that are often encountered in data preparation and pre-processing:

* Impossible or unlikely values.
* No values have been inputted.
* Irrelevant input features are present in the data at hand.

(Pintelas, 2007)

Steps are taken to eliminate the above stated problems, using various techniques and statistical methods. Statistical metrics like correlation for detecting correlated data. Correlation increase bias towards a particular feature, and also induce redundancy.

In the course of this project, different supervised machine learning algorithms are trained and tested to identify the most suitable algorithm for predicting the presence or absence of heart disease. Detailed examination of the performance of each algorithm is documented and compared to each other for determining the “best” for the problem domain.

1.5 SCOPE AND LIMITATION

In machine learning, prediction accuracy is the fundamental goal, this accuracy sometimes might be hard to verify. Verification is mostly done by domain experts, who have expert knowledge on the subject. There are other ways by which these verification can be accomplished without the effort of a human domain expert. In this project, only “Train test method” and “K-fold cross validation” are used. These verification methods have been explained in chapter four.

In terms of algorithm used, this research project focuses only on supervised machine learning method. Unsupervised learning and reinforcement learning are beyond the scope of this research project. The algorithms used includes the following:

* Naive Bayes Classifier (Gaussian Naive Bayes)
* Logistic Regression
* Support Vector Machines.

These algorithms were chosen based on the following criteria:

* Machine learning method: Every machine learning algorithm are specific to a particular machine learning method. Algorithms used for Supervised learning are neither suitable for Unsupervised learning nor Reinforcement learning and vice versa.
* Dataset: Datasets comes in different formats, some have only one observation attribute column and one observation result column. This type of dataset require algorithms like linear regression. Some datasets have multiple observation attributes that might be labeled or not, each type require algorithms that work with it’s kind of dataset.
* Result type: The result of a machine learning prediction can be in various forms, which are pre-determined by the initial form the observed result was collected. Result type can be divided into two categories namely: Regression and Classification.

Regression have a continuous set of values, e.g. The price of a car. While classification has a discrete set of values e.g. true or false, small medium or large.

1.6 DEFINITION OF TERMS

1. Machine learning: Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.
2. Dataset: a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.
3. Algorithms: A set of outlined rule to be followed in mathematical calculations or other problem solving operations like programming.
4. Observation attributes: features of a dataset that represent the characteristics of each observed unit.
5. Observation result: derived or definitive attribute of an observed unit that represent the state of the unit.
6. Artificial Intelligence: is a branch of computer science concerned with making computers act like humans and make decisions in their own.
7. Heart Disease: Diseases that affect the heart, e.g. Rheumatic heart disease, Hypertensive heart disease, Ischemic heart disease etc.

1.7 CHAPTER LAYOUT

This project document, is organized into chapters, numbered from one(1) to five(5). Each chapter describes a fundamental concept of the research process including implementation procedures.

Chapter one: This chapter lays the ground work for the whole research project. Introduction to the chapter, establishes and explains the basic concepts, the project focuses on. Other parts of the chapter, emphasizes on the specifics of the problem definition, aim and objective, Research justification and scope and limitations of the research project.

Chapter two:

Chapter three:

Chapter four:

Chapter five:

CHAPTER TWO

LITERATURE REVIEW

INTRODUCTION

BRIEF HISTORY

Because an algorithm works for one dataset does not mean it will work well with other dataset.

REVIEW OF RELEVANT LITERATURE

Relevant studies

Various publications by the pioneers of machine learning have really enhanced the way we look at machine learning. (reference publications by those who have.)

PROPOSED CONTRIBUTION

CHAPTER THREE

SYSTEM DESIGN AND METHODOLOGY

INTRODUCTION

REQUIREMENTS SPECIFICATION

SYSTEM DESIGN

Logical Design

Input/Output

Use case Diagram

Use cases

Activity Diagram

Class Diagram

Physical Design

Program Specification

Layout of Tables and Database Structure

System Controls

CHAPTER FOUR

SYSTEM IMPLEMENTATION

INTRODUCTION

IMPLEMENTATION LANGUAGE FEATURES

SYSTEM TESTING STRATEGIES

Unit Test

Integration Test

TARGET COMPUTER REQUIREMENTS

SOFTWARE MAINTENANCE ISSUES

Corrective

Preventive

Adaptive

CHAPTER FIVE

SUMMARY AND CONCLUSION

INTRODUCTION

SUMMARY

RECOMMENDATIONS (FUTURE RESEARCH)

CONCLUSION

Reference:

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.