

Republic of Iraq  
Ministry of Higher Education & Scientific Research  
University of Karbala  
College of Engineering  
Electrical & Electronic Engineering Department



---

# Natural Language Processing and Machine Learning in Medical Reports' Analysis

---

A thesis Submitted to the Department of Electrical and Electronic Engineering,  
University of Karbala in Partial Fulfillment of the Requirements for the Master's  
Degree of Science (M.sc) in Electrical and Electronic Engineering

By

**Hasanain Abdul-Jawad Hussain Almuhana**

B.Sc. in Electrical Engineering / University of Babylon

Supervised By

**Prof. Dr. Hawraa Hassan Abbas**

Muharram - 1444

August - 2022

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَقُلْ رَبِّيَ زِدْنِي عِلْمًا

صَدَقَ اللَّهُ الْعَلِيُّ الْعَظِيمُ

سورة طه الآية ١١٤

## Declaration

I hereby declare that this dissertation entitled “**Natural Language Processing and Machine Learning in Medical Reports’ Analysis**”, submitted to University of Karbala in partial fulfillment of requirements for the degree of Master in Electrical Engineering, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose sources are appropriately cited in the references.

Signature:



Name: Hasanain Abdul-Jawad Hussain

Date: 2022 / /

# Dedication

This work is dedicated to:

**My father,**

You have always supported me in every step of my life, confident of my ability to succeed. Thank you very much. I hope you are proud of your son.

**My mother,**

I know that I would not have been able to overcome these difficulties without your prayers to God to help me overcome difficulties day and night throughout my studies.

**My wife,**

Thank you very much for your support and trust that I can do it.

**my son & daughter,**

I will work hard to make up for the time I was busy studying for not being with

you.

and

**My beloved brothers and sisters**

who stands by me when things look very difficult.

# Acknowledgement

Foremost, I am highly grateful to Allah for unlimited blessings that continue to flow into my life, and because of this, I made this through against all odds.

To my supervisor, prof. **Dr. Hawraa Hassan Abbas**: I feel highly indebted to you. I am deeply grateful for your suggestions on this topic, and without your support, comments, and guidance, it would be difficult to finish this work.

To my wonderful Parents: Words cannot express my appreciation to you, the best father and mother. I wish to give you all thanks and love for your guidance, advice, invitations, and endless support.

Thank you to my friend **Hussain Ali Tuama** for the endless support. I find in my heart nothing but gratitude for what you have given me throughout my study. Thank you very much.

Furthermore, many thanks to my wonderful wife, Dr. **Zahraa Ali Tuama**, you are supportive and helpful at every stage of this thesis. My dear son **Yazan** and daughter, **Jwana**: I am sorry for being busy and did not spend much long time with you.

To my family, and my friends: I give you all thanks for your belief in me, and your constant support, encouragement, and cooperation at all times.

Last but not least, I would like to thank all the kind, helpful and lovely people who helped me directly or indirectly to complete this work and apologize to them for not being able to mention them by name here, but they are in my heart.

## Supervisor Certificate

I certify that this thesis entitled “**Natural Language Processing and Machine Learning in Medical Reports’ Analysis**” which is prepared by “**Hasanain Abdul-Jawad Hussain Almuhana**”, is under my supervision at University of Karbala in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Electronic Engineering.

Signature:



Prof. Dr. Hawraa Hassan Abbas

(Supervisor)

Date: 2022 / 10 / 6

## **Linguistic Certification**

I certify that this thesis entitled “**Natural Language Processing and Machine Learning in Medical Reports’ Analysis**” which is prepared by “**Hasanain Abdul-Jawad Hussain Almuhana**” under my linguistic supervision. It was amended to meet the English style.


Signature: *Ali J. Mahdi*

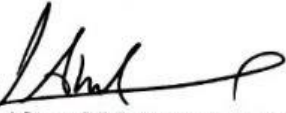
Name: **Prof. Dr. Ali Jafar Mahdi**

Date: 2022 / 10 / 6


## Certification of the Examination Committee

We, the undersigned, certify that (**Hasanain Abdul-Jawad Hussain**) candidate for the degree of Master in electrical engineering, has presented his thesis of the following title (**Natural Language Processing and Machine Learning in Medical Reports' Analysis**) as it appears on the title page and front cover of the thesis that the said thesis is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on ( 2022 / 9 / 8).


Signature:   
Name: **Dr. Hawraa Hassan Abbas**  
Title: **Professor**  
Date: 2022 / 10 / 6  
(Member) supervisor

Signature:   
Name: **Dr. Ahmed Mohammed Ahmed Alkhazzar**  
Title: **Lecturer**  
Date: 2022 / 10 / 6  
(Member)


Signature:   
Name: **Dr. Haider Galil Kamil**  
Title: **Assist. Professor**  
Date: 2022 / 10 / 6  
(Member)

Signature:   
Name **Dr. Ahmed Saleem Abbas**  
Title: **Assist. Professor**  
Date: 2022 / 10 / 6  
(Chairman) // Main

**Approval of the department of electrical and electronic engineering**

Signature:   
Name: **Prof. Dr. Haidar Ismael Shahadi**  
Title: **Professor**  
Date: 2022 / 10 / 12  
(Head of electrical department)

**Approval of deanery of collage of Engineering / University of Karbala**

Signature:   
Name: **Prof. Dr. Laith sh. Rasheed**  
Title: **Professor**  
Date: 2022 / /  
(The Dean of the College of Engineering)



# List of Abbreviations

NLP	Natural Language Processing
EMR	Electronic Medical Record
EHR	Electronic Health Record
FE	Feature Extraction
FS	Feature Selection
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
WTF-IDF	Weight and Term Frequency-Inverse Document Frequency
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
ANN	Artificial Neural Network
DNN	Deep Neural Networks
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
PCA	Principal Component Analysis
NIS	National Inpatient Sample Dataset
MIMIC	Medical Information Mart for Intensive Care
RNN	Recurrent Neural Network

MLP	Multi-Layer Perceptron
TCM	Traditional Chinese Medicine
BOW	Bag of Word
SMO	Sequential Minimal Optimization
BP	Backpropagation
DT	Decision Tree
RF	Random Forest
LSVC	Linear Support Vector Classifier
NLTK	Natural Language Toolkit
TPU	Tensor Processing Unit
GPU	Graphics Processing Unit
API	Application Programming Interface

# ABSTRACT

Natural language processing is a part of artificial intelligence algorithms that focus on designing and building applications and systems in a way that allows interaction between computers and natural languages developed for human use. NLP has been used in several areas within artificial intelligence and data processing applications. It had a positive effect on improving data quality.

The hospitals have many medical reports which contain very rich information that is not invested properly because it is unstructured data by using NLP these data transformed into structured data that can be more useful in diseases diagnosis and treatment for both the physicians and the patients.

This thesis consists of three models, that are proposed to benefit from the medical data which available in the form of text files and medical reports.

The first model is an asthma diagnosis; in this model, a semi-structured database of young patients is used. The proposed model consists of four major stages, Data collection, Data pre-processing by applying different (NLP) algorithms, features extraction and weighting by applying Weight and Term Frequency-Inverse Document Frequency (WTF-IDF) approach and finally applying the classification algorithms. The result showed that the highest accuracy (99.89%) and (97.51%) were achieved by applying ML-Perceptron on Grenada dataset and the Iraqi dataset respectively.

The second model was the classification of specialties in textual medical reports, feature extraction and feature selection methods were also used to convert the textual medical reports to sets of features and to extract the most effective feature.

Various classification methods were applied to classify dataset; the highest accuracy was achieved by applying Multi-Layer Perceptron classification techniques (99.39%).

The last model applies a deep learning algorithm on the same textual medical report's dataset that have been used in the previously mentioned model, by Applying NLP to clean data and Convolution Neural Network (CNN) which has five layers, after applying all these layers to the data in order to classify the medical reports into ten classes the result was higher accuracy equal to (99.00), F1-Measure (97.82), precession (98.64), and Recall (97.11).

## Declaration Associated with this Thesis

Some of the works presented in this thesis have been published. Appendix A refers to the paper that has been published.

- 1- **Title:** [Classification of specialties in textual medical reports based on natural language processing and feature selection.](#)

**Author:** Hasanain Abdul-Jawad Almuhana, Hawraa Hassan Abbas.

**Journal:** Indonesian Journal of Electrical Engineering and Computer Science.

**Volume:** 27,

**Number:** 1,

**Year:** 2022,

**Link:** <http://ijeecs.iaescore.com/index.php/IJECS/article/view/27023/16464>

- 2- **Title:** [Classification of Specialties in Textual Medical Reports Based on Natural Language Processing and Deep Learning](#)

**Author:** Hasanain Abdul-Jawad Almuhana, Hawraa Hassan Abbas.

**Journal:** International Journal of Health Science.

**Volume:** 27,

**Number:** 1,

**Year:** 2022,

**Link:** <https://sciencescholar.us/journal/index.php/ijhs/article/view/12476>

- 3- [Asthma Classification Based on Natural Language Processing and Machine Learning](#)

This paper was accepted from international conference of engineering sciences (**ICES 2021**)

and will be publish soon.

# List of Contents

Declaration.....	III
Dedication .....	IV
Acknowledgment.....	V
Supervisor Certificate .....	VI
Linguistic Certification .....	VII
Certification of The Examination Committee.....	VIII
List of Abbreviations .....	IX
Abstract .....	XI
Declaration Associated With This Thesis.....	XIII
Table of Content .....	XIV
List of Figures.....	XX
List of Tables .....	XXIII
<b>CHAPTER ONE Introduction.....</b>	<b>1</b>
1.1 Overview .....	1
1.2 Problem Statement .....	4
1.3 Aim of Thesis.....	4
1.4 Thesis Objectives .....	5
1.5 Thesis Outlines: .....	6

<b>Chapter 2</b>	<b>Literature Review:.....</b>	<b>7</b>
2.1	NLP in Automatic Diseases Diagnosing.....	7
2.2	Classification Medical Text Reports.....	10
2.2.1	Traditional Feature Extraction and Machine Learning Techniques...10	
2.2.2	Deep Learning Techniques.....	15
<b>Chapter 3</b>	<b>Theoretical Background .....</b>	<b>19</b>
3.1	Introduction .....	19
3.2	Electronic Medical Records (EMRs).....	19
3.3	Natural Language Processing (NLP). .....	20
3.3.1	The Power of NLP.....	21
3.4	Text Pre-Processing .....	22
3.5	Feature Extraction (FE).....	23
3.5.1	Bag Of Words (BOW). .....	23
3.5.2	Term Frequency-Inverse Document Frequency (TF-IDF). .....	24
3.6	Feature Selection (FS).....	25
3.6.1	Chi-Square Test. ....	26
3.6.2	Principal Component Analysis (PCA).....	27
3.7	Machine Learning .....	28
3.7.1	Text Mining (TM). .....	29
3.7.1.1	Support Vector Machine (SVM). ....	30
3.7.1.2	Multiple Layer Perceptron (MLP). ....	32
3.7.1.3	Decision Tree (DT).....	33

3.7.2 Deep Learning.....	34
3.7.3 One-Hot Full Embedding .....	36
3.7.4 Word Embeddings. ....	36
3.7.5 Convolutional Neural Networks CNNs. ....	38
3.7.5.1 Convolution Layer. ....	39
3.7.5.2 Pooling Layer. ....	40
3.7.5.3 Flatten Layer. ....	41
3.7.5.4 Fully Connected Layer. ....	41
3.7.5.5 Dense Layer. ....	42
3.7.6 Parameters.....	43
3.8 Performance Metric For Classification Algorithms.. ....	46
3.8.1 Accuracy.....	47
3.8.2 Precision.. ....	47
3.8.3 Recall. ....	48
3.8.4 . F-Measure.. ....	48
3.8.5 Cross Validation.....	49
<b>Chapter 4 Methodology.....</b>	<b>50</b>
4.1 Introduction.....	50
4.2 The Proposed System Architecture.....	51
4.2.1 Asthma Diagnoses Model.....	51
4.2.2 Proposed System for Classical Classifier Model. ....	52



4.2.2 Deep Learning Classifier Model.....	54
4.3 Datasets.....	55
4.3.1 Asthma In Childhood (ISAAC) Dataset Grenada University..	57
4.3.2 Asthma Data Set From Iraqi Hospitals. ....	60
4.3.3 Emrs Patient Medical Reports Dataset.....	63
4.4 Asthma Diagnoses Using NLP & Data Mining. ....	64
4.4.1 Pre-Processing Stage. ....	65
4.4.2 Feature Extraction and Weighting Stage.....	67
4.4.2.1 Feature Extraction. ....	67
4.4.2.2 Feature Weighting. ....	68
4.4.3 Text Mining Stage.....	69
4.5 Classification Of Specialities In Textial Medical Reports By Using Classical Algorithms. ....	70
4.5.1 Data Pre-Processing: .....	70
4.5.1.1 Word Tokenization.....	71
4.5.1.2 Cleaning Data.....	71
4.5.1.3 Removing Punctuations.....	71
4.5.1.4 Removing Symbols And Special Characters. ....	72
4.5.1.5 Removing Non-English Letters. ....	72
4.5.1.6 Removing Stop Words. ....	72
4.5.1.7 Conversion to Lower Case. ....	72
4.5.2 Feature Extraction (FE).....	73

4.5.2.1 Creating A Vector Of Features. ....	73
4.5.2.2 Calculating Weights For Features.....	73
4.5.3 Feature Selection (Fs) By Using Chi-Square Test. ....	74
4.5.4 Data Classification. ....	74
4.6 Classification of Specialities In Textual Medical Reports By Using Deep Learning. ....	75
4.6.1 Data Pre-Processing. ....	75
4.6.1.1 Clean Data.....	75
4.6.1.2 Encoding Techniques (Vector Representations of Text)...	76
4.6.2 Deep Learning Layers. ....	76
4.6.2.1 Word Embedded Layer. ....	76
4.6.2.2 Convolution Neural Network Layer. ....	77
4.6.2.3 Pooling Layer.....	77
4.6.2.4 Flatten Layer. ....	78
4.6.2.5 Dense Layer. ....	78
4.6.2.6 The Fully-Connected Layer.....	78
4.6.3 libraries used in DL model .....	79
4.6.3.1 The TensorFlow .....	79
4.6.3.2 Keras.....	79

<b>Chapter 5</b>	<b>Results and Discussion .....</b>	<b>80</b>
5.1	Introduction .....	80
5.2	Software and Hardware.....	80
5.3	Result of Asthma Diagnoses Model. ....	81
5.3.1	Prepossessing Result... ..	81
5.3.2	Feature Extraction Result... ..	82
5.3.3	Classifier Result... ..	83
5.4	Results of Classical Classification Model. ....	92
5.4.1	Results of Data Pre-Processing. ....	92
5.4.2	Results of Features Extraction.....	94
5.4.3	Results of The Classification Algorithms. ....	94
5.5	Result of Deep Learning CNN Classifier. ....	97
5.5.1	Preprocessing Results.....	97
5.5.2	Embedding Results.....	97
5.5.3	CNN Results... ..	97
5.6	Discussion.....	99
<b>CHAPTER 6</b>	<b>Conclusion &amp; Future Work.....</b>	<b>102</b>
6.1	Conclusion .....	102
6.2	Future Works .....	104
<b>References</b>	<b>.....</b>	<b>105</b>

## List of Figures

Figure (3:1) Electronic Medical Records .....	20
Figure (3:2) Feature Selection .....	25
Figure( 3:3 ) Principal Component Analysis .....	27
Figure( 3:4 ) Illustration of Hyperplane.....	30
Figure (3:5) Support Vector Machinefigure .....	31
Figure (3:6) Multiple Layer Perceptron (MLP) .....	32
Figure (3.7) Decision Tree .....	33
Figure (3:8) Simple of Convolution Neural Networks Layers.....	35
Figure (3:9) The Relationship Between Acc. and Amount of Data.....	35
Figure (3:10) Coding Diagram of One-Hot Encoding and WE Encoding ....	37
Figure (3:11) Simple of Cnn For Text Document .....	38
Figure (3:12) Typical Convolutional Model For Texts.....	39
Figure (3:13) Simple of Pooling Layer .....	40
Figure (3:14) Embedding Layer and Flatten Layer.....	41
Figure (3:15) Flatten Layer and Fully Connected Layer.....	42
Figure (3:16) Kernel .....	43
Figure (3:17) Stride .....	43
Figure (3:18) Padding .....	43
Figure (3:19) Bias .....	44
Figure (3:20) Intuition .....	44
Figure (3:21) Filters.....	45

Figure (3:22) 10-Fold Cross Validation.....	49
Figure (4:1) Asthma Diagnoses Model.....	52
Figure (4:2) Proposed System For Classical Classifier Model.....	53
Figure (4: 3) Propose System of Deep Learning Classifier.....	54
Figure (4.4 A) Patient Survey for Grenada Dataset.....	58
Figure (4.4 B) Patient Survey for Grenada Dataset.....	59
Figure (4.5) A Iraqi Patient Survey. ....	61
Figure (4.5) B Iraqi Patient Survey. ....	62
Figure (4:6) Sample of Emr Dataset.....	63
Figure (4:7) Pre-Processing Steps.....	64
Figure (4:8) Block Diagram of the Pre-Processing Steps... ..	70
Figure (5:1) Sample of Data After Preprocessing... ..,	81
Figure (5:2) File of Medication Column After Feature Extraction.....	82
Figure (5:3) File of Relative Column After Feature Extraction .....	82
Figure (5:4) File of Structure Column's.....	82
Figure (5:5) The Confusion Matrix of The Three Algorithms For Structured Grenada Dataset .....	84
Figure (5:6) The Confusion Matrix of The Three Algorithms For Semi Structured Grenada Dataset Without Fs Technique.....	86
Figure (5:7) The Confusion Matrix of The Three Algorithms For Semi Structured Grenada Dataset With Fs Technique.....	87
Figure (5:8) The Confusion Matrix of The Three Algorithms For Semi Structured Iraqi Dataset With Fs Technique .....	88

Figure (5:9) The Confusion Matrix of The Three Algorithms For Semi Structured Grenada Dataset With Fs Technique.....91

Figure (5:10) The Confusion Matrix of The Three Algorithms With Chi - Square Technique ..... 96

Figure (5.11) Deep Learning CNN Results.....98

## List of Tables

Table (3.1) Confusion Matrix.....	46
Table (4:1) Sample of The Emrs Dataset.....	56
Table 4:2 The Distribution of Dataset. ....	56
Table (4:3) Number of Instances For Each Medical Specialty .....	57
Table (5:1) The Performance Metric of The Three Algorithms With NLP ....	83
Table (5:2) The Performance Metric of 3 Algorithms Without FS Technique.	85
Table (5:3) The Performance Metric With Chi- FS Technique .....	87
Table (5: 4) The Performance Metric without NLP for Iraqi Dataset.....	89
Table (5:5) The Performance Metric of The Three Algorithms With NLP Technique For Iraqi Dataset and Without FS (10 Cross Validation) .....	89
Table (5:6) The Performance Metric of the Three Algorithms With NLP Technique for Iraqi Dataset and with FS (10 Cross Validation) .....	90
Table (5:7) Word Tokenization.....	92
Table (5:8) Removing Punctuation.....	93
Table (5.9) The Result of Preprocess Step on The Text of The EMR .....	93
Table (5:10) The Performance Metric of The Three Algorithms without FS...94	
Table (5:11) The Performance Metric with FS .....	95
Table (5:12) The Execution Time of The Three Algorithms.....	95
Table (5:13) Deep Learning Results.....	97
Table (5:14) Comparison of Classical Classification DL CNN Results.....	98
Table (5:15) Methods and Result For Classification Related Work .....	100

# CHAPTER 1

## INTRODUCTION

### 1.1 Overview

Programming is essential development in the medical field, especially in data mining, to obtain information and classifications based on the basic principles of medical data [1]. The medical field contains many valuable and useful information which, if properly organized, can be more helpful in diseases diagnosis and treatment [2].

There is huge interest in applying artificial intelligence (AI) in the medical field to improve diagnosis methods which can be applied in public health, patient care, and pharmaceutical researches. The success of AI systems in such analysis depends on the availability of datasets and quality of the included data, however, pre-processing unstructured data is an important step towards acquiring efficient information from medical records using AI techniques [3].

NLP is a branch of AI, aims at primarily reducing the distance between the capabilities of a human and a machine and analyse the natural information data such as images, text documents and any unstructured data into structured data can be processed. NLP, AI, and ML all have the potential to simplify the use of unstructured data [4].

Programming companies have identified several more applications of NLP in healthcare, to achieve two important targets, administrative cost reduction and medical value creation [5].



Asthma is a reversible airway obstruction that leads to narrowing of the bronchial lumen with hypertrophy of the bronchial wall and over secretory mucous gland. For some asthmatic patients, who are mildly affected, asthma is a simple disease and is not very disturbing, while for others who are severely affected, it can be very harmful and it is a problem that interferes with daily activities of life such as walking or normal movement and may lead to a life-threatening asthma attack. Asthma cannot be cured, but its symptoms can be controlled, this is because asthma often changes in severity over time from one level to another, so it is important to keep track of your signs and symptoms with your doctor and adjust treatment from time to time as needed [3]. Asthma symptoms vary from one person to another and it has multiple symptoms, the most important one is shortness of breath, wheezing and chest pain or tightness in the chest [4]. Asthma is a common disease in Iraq as a result of air pollution with several pollutants. Consequently, automatic diagnosing and classification of this disease is an important task for establishing an automatic health care system in Iraq [6].

Most of the hospitals around the world have modern health information systems such as Electronic Health Records (EHRs), electronic medical records (EMRs), and other systems. EMR includes the patient's full medical and treatment history in a single file.

Digital healthcare data was estimated to be equal to 500 petabytes Worldwide and is expected to reach 30 Exabyte in 2022. Predicted that the global growth in healthcare data will be between 1.2 and 2.4 Exabyte [7].

NLP analysis to examine social media and medical reports data is an effective way to improve treatments and patient services by understanding how patients and the medical staff talk about treatments, drugs, and diet practices [5].

With NLP and text mining, healthcare organizations are starting to leverage technology to access the plethora of unstructured patient data available in the EMR (e.g., nursing notes or patient-reported text). NLP and text mining can process data that traditional analysis cannot process and thus opening up richer and more complex data sources [8].

Traditional analytics typically uses structured data, consisting mainly of claims data and only accounting for approximately 20 percent of all available data. Structured data exists in a specific, consistent format and includes basic information, but not valuable details. Unstructured data include free text data, physician order data, nursing notes, and dictation notes. To access the remaining 80 % of this data, one should rely on NLP. That allows organizations to access the complex, richer data sets that are harder to reach because they require sophisticated technology to derive value from the massive amounts of everyday language setting in the EHR [8].

Unstructured data, residing in EHRs and elsewhere, contains the deeper, more complex information such as a patient's own words to describe symptoms or doctors' notes.

In this work, three models based on NLP are proposed,

A diagnosing model for diagnosing asthma was generated from semi-structured data containing a structured part and an unstructured part of the data. The second model was a classifier that classified the text medical reports into nine categories by using traditional algorithms.

The proposed classifier is improved by using deep learning technique and it is used to process the same data in the second model to get better results.

## **1.2 Problem Statement**

While medical data is growing exponentially, a bulk of it remains unused, mainly because healthcare-related information systems are not equipped to process unstructured data. If the capabilities to interpret, analyze and utilize this unstructured medical data were available, the benefits would be tremendous both for patient treatment, as well as for public health management and medical research.

For clinical research, a huge quantity of detailed patient information, such as disease status, side effects, medication history, treatment outcomes, and lab tests, has been collected in an electronic format called electronic medical record (EMR) and it serves as a valuable data source for further analysis. Therefore, a huge quantity of detailed patient information is present in the medical text, and it is quite a huge challenge to process it efficiently.

## **1.3 Aim of thesis**

The aim of this thesis is developing an automatic classification and diagnosing systems, which specializes in extracting the maximum useful information from medical datasets.

The medical dataset contains structured data, semi-structured data and unstructured data. All mining methods were extracting information from only structured data. The major part of this thesis contains preprocesses for data to be able to process from algorithms to get the maximum amount of information from the medical datasets, by applying some AI techniques that led to the best accuracy in classification such as feature extraction and feature selection techniques.

## **1.4 Thesis Objectives**

The objectives can be summated as follows:

1. Proposing a mathematical model based on NLP for diagnosing asthma was generated from semi-structured data which contain a structured part and an unstructured part of the data.
2. Developing a mathematical model based on NLP to design a classifier which classifies the text written in medical reports into nine categories by using traditional algorithms.
3. Design a classifier model that works on the principle of deep learning technique and it is used to process the text medical reports into ten categories.

## **1.5 Thesis Outlines**

This thesis is divided into six chapters. Each chapter begins with a short overview that offers a general impression on the chapter.

The thesis's structure is organized as follows:

- In chapter 2, brief historical literature of asthma diagnosis techniques, including traditional diagnosis methods. Then, the Medical Reports classification with classical methods will be reviewed, advancing finally to the relevant modern employment of DL in text classification for medical reports.
- In chapter 3, the theoretical background of this thesis will be discussed in details, including a brief overview of Asthma disease, its causes, symptoms, diagnosis and medical text reports. This chapter also provides a detailed explanation of NLP, FE, FS, AI and DL techniques and their uses in computer aid classification and diagnosis technology.
- In chapter 4, a full presentation of the datasets used in this thesis and the proposed systems for each model and all techniques used in this work will be described step by step.

- In chapter 5, a demonstration of the results of the proposed system and the research experiments. It also discusses the evaluation of the system's performance.
- In chapter 6, thesis conclusions and recommendations for future work are detailed.

# **CHAPTER 2**

## **LITERATURE REVIEW**

In this chapter, a brief overview is presented of the studies that discuss how to automatically extract information from different types of electronic medical records for the purposes of disease diagnosis "Asthma", as well as for classification purposes, using classical algorithms, classification techniques. First, the classification of asthma for semi-structured datasets is discussed. Then, traditional feature extraction, feature selection, and classification methods based on classical ML techniques for unstructured data are discussed in detail. Finally explore the different DL models used in EMRs classification and a summary of previous methods is provided.

### **2.1 NLP in Automatic Diseases Diagnosis**

NLP is essential for improving the healthcare services and disease diagnosis by interpreting clinical notes effectively. It extracts details from patients' Medical reports and doctors' letters and ensures the completeness and accuracy of patient health profiles.

Asthma is a well-known disease in the world and affects the respiratory system of many people in different age groups [9]. Many medical research centers have been established to study, diagnose and treat asthma as a common disease ,in spite of the 262 million people in the world suffer from Asthma disease, little number of researches is conducted to diagnose it automatically [10] .

(Mathur and Joshi [11], 2019), the researchers proposed a model for automatic diagnosing of asthma in children using AI by applying the model to a semi-structured database from the University of Grenada “**the same database that was used in the diagnostic model in this thesis**”, by used a Naïve Bayesian (NB) classification technique, that works on the basis of the theory of probability, without any NLP pre-processing tools

and compared algorithms on the basis of three Measures, Recall, Precision, and F-Measure. without specifying the percentages of accuracy and efficiency.

(Kukreja, and Saksham [12], 2018), They focused on developing and evaluating algorithms for diagnosing asthma based on questionnaires collected via the mobile application on Android and iOS that asks questions about severe symptoms, patients' medical reports, and clinical data for patient asthma. ML algorithms, including back propagation model, C4.5 algorithms, and Bayesian networks, were used. All algorithms received more than 80% accuracy and found that the death rate from asthma could be reduced by 78% if the patient was close to the nebulizer device.

(Wu, Stephen T., et al. [13], 2013), developed (NLP) system to extract predetermined criteria for asthma from unstructured text in EMRs using manual chart reviews as a gold standard, asthma status (yes vs no), and identification date (first date of a “yes” asthma status) were determined by the NLP system. The result showed patients were a group of children (n = 112) 84% Caucasian, (49%) girls younger than 4 years (mean 2.0 years) who participated in previous studies. The NLP approach to asthma ascertainment showed sensitivity, specificity, positive predictive value, negative predictive value, and median delay in diagnosis of (84.6%), (96.5%), (88.0%), (95.4%), respectively.

Similarly, (Do et al. [14], 2017), tried to take advantage of artificial neural network machine learning to identify cases of asthma. They conducted research on international inpatient databases (national inpatient sample, NIS) and hospital-level databases (MIMIC III). The sample size for the NIS database is 68,847 asthmatic patients while MIMIC III database has 214 patients.

(Wi et al. [15], 2017), Thus tried to take the advantage of NLP in the process of identifying asthma. They assessed NLP performance on a database containing 500 samples, producing (sensitivity, specificity, positive predictive values, and negative predictive values) of (97%, 95%, 90%, and 98%) respectively.

Similar works' problems in Asthma diagnosis can be tackled by the following points.

- 1- Some algorithms have achieved good percentages in Recall, Precision and F-Measure. But at the same time, they achieved low accuracy rates [11] .
- 2- Survey data via mobile applications is highly unreliable [12] .
- 3- The percentage of accuracy achieved is acceptable but the adoption of the model is insufficient for the diagnosis [12] .
- 4- Some researchers needed a relatively large database to reach an acceptable accuracy level that exceeds 65,000 samples. This leads to a slow execution time and the need for advanced processors [14].

The gap in disease diagnosis researches was that most of the researches reached a somewhat acceptable accuracy, (between 70% to 90 %) as a maximum, and was often directly proportional to the size of the data, but it was not suitable for diagnosis because it is related to people's lives and does not accept an error rate, no matter how small it was. One of the goals of this thesis was to reach a high accuracy rate using a database relatively simple and uncomplicated.



## **2.2 Classification of The Medical Text Reports.**

### **2.2.1 Traditional Feature Extraction and Machine Learning Techniques.**

Automatic medical text classification is highly useful in the (NLP). For medical text classification tasks, ML techniques seem to be quite effective; however, it requires extensive effort from the human side, so that the labelled training data can be created. For clinical and translational research, a huge quantity of detailed patient information, such as medication history, disease status, lab tests, side effects, and treatment outcomes has been collected in an electronic format called EMRs, and it serves as a valuable data source for further analysis. therefore, a huge quantity of detailed patient information is present in the medical text, and it is quite a big challenge to process it efficiently [16].

EMRs are increasingly being used to document and store large amounts of clinical information, however efficiently and accurately extracting meaningful data from electronic health records is a challenge, as a significant portion of clinical information is stored in unstructured text. Electronic medical records (EMR) provide unique possibilities for clinical research; However, some important patient traits are not readily available due to their unstructured properties because these data are saved in a format that algorithms cannot process. Hence, manual review of these data is often time-consuming, error-prone and expensive [17].

As most medical data contain long diagnostic texts and medical reports that may contain repetitive and varied medical terms or words that algorithms cannot handle, the application of NLP has been found to be necessary for medical data processing to obtain the maximum amount of usable information [18].

ML algorithms play a vital role in the medical field automatic diagnosing system establishing models to help researchers, technologists, and clinicians to create more objective systems that provide support to objectify the diagnosis, save resources, and improve treatment efficacy [19].

(Lucini, Filipe R., et al. [20], 2017), Tested a different approaches for pre-processing text records and to predict hospitalization. Sets-of-words are obtained via binary representation, term frequency, and other approaches are tested for feature formation, then used feature selection to decrease the decimation of data. They used eight methods for tested text mining: Support Vector Machine (Kernel linear) Random Forest, Extremely Randomized Tree, Multinomial Naïve Bayes, Decision Tree, AdaBoost, Logistic Regression, and Nu-Support Vector Machine (Kernel linear). The best acquired results were (77.70%) with Nu-Support Vector Machine.

(mXie, Jingui, et al. [21], 2020), worked on understand the nature of diseases and to improve the treatment of Traditional Chinese Medicine (TCM) syndromes and select critical features of demographic information, personal medical history, and symptoms, and improve the accuracy of syndrome classification by collecting a total of 1713 records from Hospital of Anhui Chinese Medicine University. Five rules for feature selection and six models were applied to classify TCM syndromes.

In total, 200 features were extracted from electronic medical records, 42 were selected as critical features. The classification accuracy of using feature selection was higher than when using all features, with a maximum value of 0.88 for the Artificial neural network (ANN).

(Li et al. [22], 2021), Prepared a review study focused on a broad scope of AI medical tasks, such as classification and prediction, word embedding, feature extraction, feature generation, and similar matters such as question answering, phenotyping, generating knowledge graphs, forming a medical dialogue, and supporting multilingual communication and interpretability, and reviewed multiple recent studies that showed how such tasks could be supported by electronic health records and health informatics, concluding that Deep learning methods in the general field of NLP have achieved remarkable success, but that applying them to the field of biomedicine remains challenging due to limited data availability and additional difficulties associated with domain-specific text data.

(Soguero-Ruiz, Cristina, et al. [23], 2014), Designed a model of early detection of severe complication after elective surgery for colorectal cancer surgery called anastomosis leakage, using text documents extracted from EHRs. By using a bag of words model to investigate the potential for feature selection strategies. The purpose is earlier detection of this leakage and prediction of it with data generated in the EHR before the actual complication occurs.

In that paper, the feature selection strategies are used to decrease the high dimensionality of the data, for mining they derived the robust support vector machine linear maximum margin classifier for three classes. Results reported a discriminatory power for early detection of complications after colorectal cancer surgery (sensitivity 100%; specificity 72%).

(Alshaer, Hadeel N., et al. [24], 2021), Studied the effect of the improved chi Square method on the performance of six well-known classifiers, Random Forest, Decision Tree, Naïve Bayes, Naïve Bayes Multinomial, Bayes Net, and Artificial Neural Networks. The dataset employed in this paper includes 9055 Arabic documents that were collected from various Arabic resources. Based on their content, these documents were divided into twelve categories.

(Caccamisi et al. [25], 2020), Developed a model using machine-learning algorithms to automatically classify patients' smoking habits. This applied text mining, using machine learning to enable automatic classification of unstructured information on smoking status drawn from Swedish EMR data. Data of patient smoking status from EMRs was then used to develop 32 different predictive dataset selections across a database of 85,000 classified sentences. The best performing model was based on the Support Vector Machine (SVM), Sequential Minimal Optimization (SMO) classifier. Sentence frequency and attribute selection did not improve model performance, however. SMO achieved 98.14% accuracy and an F-score of 0.981 versus values of 79.32% and 0.756 for the rule-based model.

(De la Torre, Juan, et al. [26], 2020), Focused on a case study assessment of cervical by using (DM) and (ML) techniques to show the ability to generate reliable predictive models in the field of healthcare. Using a database of 302 samples. They have generated several predictive models, including logistic regression, SVM, k-NN, gradient boosting, DT, random forest, and neural network algorithms. the goal is to predict the potential presence of cervical pain in patients affected with whiplash diseases. The results show that it is possible to reliably predict the presence of cervical pain (accuracy, recall, and precision above 90%).

(Solti, Imre, et al. [27], 2009), were compares the performance of keyword and machine learning-based chest x-ray report classification for lung injury. the goal of this study was to create an automated system that could reduce the time needed to recognize injuries lead to reductions in mortality rate. by using 857-reports for chest x-ray reports. they were labeled by domain experts.

Word unigram and character n-grams provided the highest performance (Recall equal to 0.91, F-measure equal to 0.91, and Precision equal to 0.90).

One of the promising trends in this field is the development of better knowledge of mining information from unstructured data [28], which is useful when working with a combination of structured and unstructured data to develop better decision making and facilitate broader interpretation.

Although traditional feature extraction and machine learning techniques achieved good classification results, these still suffer from these limitations:

1. Clinical notes in a particular patient's record may contain a lot of redundancy, due to many physicians' habits in copying notes and pasting them into a new note.
2. Frequency of the word often does not indicate the importance and impact of the word when using feature selection. This leads to an error rate, especially when the classification is binary representation [20].
3. Classic classifiers need a large database for testing and training to be able to classify correctly and with high accuracy [25].
4. The accuracy decreases whenever the required classification is multiple, i.e., the classification accuracy is inversely proportional to the number of required classifiers and is directly proportional to the volume of available data.
5. Requires intensive processing steps before manual or automated feature extraction and selection.
6. Extracting low-level features using classical ML algorithms could fail to achieve the best results.
7. Traditional ML approaches have relatively lower performance with larger amounts of input data.

## 2.2.2 Deep Learning Techniques

Deep learning is a part of ML include a neural network with three layers at least. These neural networks attempt to simulate the behaviour of the human brain allowing it to “learn” from large amounts of data.

Deep learning (DL) techniques have begun to dominate because of their simplicity and no need for handcrafted features and efficient processing, meanwhile [29].

Convolutional neural networks deep learning (CNNs) has dramatically improved the approaches to solve many research problems.

One of the key differentiators between CNNs and traditional machine learning approaches is the ability of CNNs to learn complex feature representations [30].

(Fesseha, Awet, et al. [31], 2021), Studies using CNN for Tigrinya which is a Semitic language spoken in Ethiopia and Eritrea by constructing a CNN with a continuous bag-of-words feature extraction method. by developing model that has divided 30,000 text documents into six labels. (categories) include “health”, “politics” “agriculture”, “religion”, “sport” and “education”. CNN's with word2vec method, CNNs without word2vec method. The results found that The CBOW CNN with word2vec achieves the best accuracy with (93.41%).

(Geraci, Joseph, et al.[32], 2017), Provided a study of applying ML to diagnose youth depression the phenotyping from psychiatrists by using 861 labelled documents from EMRs. The goal was a model to identify individuals who meet inclusion criteria, two psychiatrists who labelled a set of 861 EMRs documents, using a brute force search and training a deep neural network and according to a cross-validation evaluation. The result showed that the model had a specificity of (97%) and a sensitivity of (45%).

(Weng et al. [33], 2017), Showed that a supervised learning-based NLP approach is useful for developing medical subdomain classifiers. The medical subdomain, such as cardiology or neurology, of a clinical note is a useful piece of content-derived metadata for developing machine learning downstream applications. Two datasets, the iDASH data repository (n = 431) and the Massachusetts General Hospital database (n = 91,237) were used.

Using Term Frequency-inverse document frequency (TF-IDF) improved the result and proved that weighting outperformed other learning classifier datasets, with an accuracy of 0.957 and 0.964 and F1 scores of 0.932 and 0.934 for the two datasets, respectively. Classifiers were trained on one dataset and applied to the other dataset, yielding a threshold F1 score of 0.7 with regard to classifiers for half of the medical subdomains.

(Mu, Xiuli, and Hongyan Zhang. [34], 2021), Constructed (Pure CNN) based on pooled and non-pooled analysis is based on deep learning CNN to mine EMRs text reports. to extract the pregnant women medical information from EMRs text reports. CNN is adopted to classify the Chinese character tags of the neural network model. The model is applied the word segmentation of the actual pregnant women's EMR text. The required results showed the strong stability of the model when it was conducted on Biomedical Engineering (Peking University) and Microsoft Reserved Partition datasets are between 0.9516 and 0.9684.

(Thomas et al. [35], 2014), Performed a study to assess the validity of an NLP program designed to accurately identify patients with prostate cancer: to achieve this, a retrospective review was performed of a prospectively collected database that featured patients from the Southern California Kaiser Permanente. A consecutive series of 18,453 pathology reports were evaluated, and NLP was found to correctly detect 117 out of 118 patients.

This translated to a positive predictive value of 99.1% with 99.1% sensitivity and 99.9% specificity in terms of correctly identify patients with prostatic adenocarcinoma after biopsy. The overall ability of the NLP application to accurately extract variables from the pathology reports was 97.6%.

([Hammoud et al.\[36\], 2021](#)), Presented a new Arabic medical dataset for text classification. The dataset included 2,000 articles invaded into 10 classes (Blood, Bone, Cardiovascular, Ear, Endocrine, Eye, Gastrointestinal, Immune, Liver, and Nephrological) of disease. The model was pre-trained on an Arabic medical reports corpus before fine-tuning on the relevant dataset with the original model produced results of (97.4331) for F1 Validation. and 95.9124 in F1 Testing, with overall SVMs of (89.1308) and (87.3473), respectively.

([Khichdee et al. \[37\], 2016](#)), Introduced an instrument for classification of medical records based on the language, based on 24,855 text records. The documents were classified into three groups (endoscopy, ultrasonography, and X-ray), with 13 subgroups, using two methods; these were Support Vector Machine (SVM) and K-Nearest Neighbour (KNN)with feature selection.

At the second stage of classification into subclasses, 23% of all documents could not be linked to only one definite individual subclass (binary system or liver) due to the common features characterising these subclasses.

The methods for creating a model depend on the structure and the characteristics of the relevant database. This model used a unique dataset to train the algorithm to classify medical text reports, which enables the resulting model to classify any text report within the medical field. Medical reports can be classified into six major classes based on results and data quality.



Clinical notes in a particular patient's record may contain a lot of redundancy, due to the large part of the documentation habit of many physicians by copying past notes and pasting them into a new notes [38].

Most studies that used DL techniques suffered from three limitations:

- 1- These techniques are time-consuming
- 2- These techniques require a great deal of collecting data to produce an accurate diagnosis.
- 3- These techniques also require intensive pre-processing steps to facilitate feature extraction and selection, which can be error-prone.

Most Recent studies depended on more advanced DL techniques.

- 1- They are computationally intensive because of CNN complexity.
- 2- They are prone to over-fitting due to high dimensionality.

## CHAPTER 3

### THEORETICAL BACKGROUND

#### 3.1 Introduction

this chapter, provide a brief overview of the types of electronic medical records, their contents, and the method for collecting the data they consist of, with a comparison of the types of electronic medical records.

NLP and its most important applications especially word processing and classification are also defined. Feature extraction and feature selection and cross-validation techniques are described. A general idea of the data quality parameter is given. The role of data mining has been described.

The classic machine learning algorithms and deep learning methods are explained.

#### 3.2 Electronic Medical Records (EMRs)

A digital copy of the paper charts seen in many clinicians' offices exists in most cases, and this is known as an electronic medical record or EMR. It was online records include of the standard clinical data and medical data from a single medical source or medical service office. As shown in Figure 3.1, an EMR includes the patient's full medical and treatment history in a single file. Thus, the main advantage of EMRs allowing clinicians to track data over time and easily identify where patients need check-ups or preventive screenings, as well as allowing them to check how well their patients are complying with treatment and performing against certain parameters such as blood pressure readings, improving the overall quality of care and service [39].

EMR used by treatment providers for diagnosis or to check a patient’s medical history. Comprehensive and accurate documentation of tests, diagnoses, and treatment in EMRs ensures appropriate care throughout a patient’s experience with a given provider [40].



*Figure 3.1 Electronic medical records [39].*

EMR are increasingly being used to document and to store large amounts of clinical information, yet efficiently and accurately extracting meaningful data from EHRs is challenging, as a significant portion of clinical information is stored in unstructured, free text. Manual review of this data is thus often necessary, which can be time-consuming, error prone, and costly [17].

### **3.3 Natural Language Processing**

(NLP) is a broad and specialized field of computational linguistics, and artificial intelligence which attempts to deal with human language [41]. NLP primarily concentrates on the design and construction of applications and systems that permit interaction among computers and natural languages which are evolved to human use. Typically, NLP techniques empower computers to process and understand natural human language and utilize it to provide useful output [42]. It is difficult to deal with natural languages directly as they contain noise. Such noise cannot be processed directly.

NLP has many tasks such as Text Generation, Text Classification, Machine Translation, Speech Recognition, Sentiment Analysis, etc. For a beginner to NLP, looking at these tasks and all the techniques involved in handling such tasks can be quite daunting. It is very difficult for a newbie to know exactly where and how to start.

### **3.3.1 The Power of NLP**

NLP is extremely powerful, primarily because language is ubiquitous and also because tools to analyze language automatically provide indices related to virtually any aspect of language. NLP can detect the specific words used, groups of words, and the strength of the relations between words and between larger bodies of text. It can also detect the features of the text, such as the frequency, concreteness, or meaningfulness of the words, the complexity of the sentences, and various aspects of the text such as cohesion and genre. The words and their features serve as proxies for various constructs. For example, the frequency of the words in a text serves as a proxy to estimate the knowledge that might be required to understand the text. The cohesion of a text affords an estimate of the knowledge necessary to fill in the gaps in text. Developing tools to calculate linguistic indices is the main part in this research, that move beyond these surface-level tasks and provide information that may be more important within educational contexts. Notably, describe a subset of NLP techniques that provide information about multiple levels of text. These tools begin from the words in the discourse, extract specific word features, and then go beyond the lexicon by considering semantics, as well as discourse structure [43]. The aim of this research is to provide cases of a few common techniques, rather than an overview of all available methods. These methods grouped into those that focus on the words directly as the units of analysis, and those that focus on features of the words.

### 3.4 Text Pre-processing

Collection of a large number of textual data is a significant process for learning classification algorithms. Text pre-processing is the method that converts texts into an appropriate form to be classified. Text pre-processing is very important because it can be used to decrease the feature space and computational process, and this, in turn, can positively affect the classification accuracy. Many pre-processing steps are performed to prepare texts for classification. For example, all unnecessary data such as non-English words, special characters, symbols, usernames, repeated and elongated letters, numbers, digits, and punctuations are removed. Unnecessary information that often reduces the efficiency and accuracy of the classification algorithm can also be removed [44]. Most important preprocessing steps are discussed here, word tokenization is the method of breaking up or segmenting sentences into their component words. The reason for that is to reconstruct words sentence[45]. Tokenization is very important in many processing steps to split the text document into tokens based on space, commas, periods, semicolons, and quotes between words. These tokens may be symbols, words and /or numbers.

Typically, Python library provides various interfaces for performing word tokenization[46]. Another task in the pre-processing phase is to eliminate unimportant words such as stop words. Stop words are words that do not have any significance or useful information in the text such as prepositions, pronouns, and conjunctions. They are usually removed from the text during processing in order to reduce the size of the text to be processed by the classification algorithm, whereas the retaining words have the maximum significance and context. Stop words are usually words which occur frequently in the vocabulary of a language. There is no comprehensive list of stop words, so each language has its own set of stop words[47].

### **3.5 Feature Extraction**

(FE) process is performed in order to extract meaningful features or attributes from a raw textual in order to be prepared for feeding to a statistical or machine learning methods. This process is known as vectorization because the result of this process is numerical vectors from raw text tokens. The reason is that machine learning algorithms work on numerical vectors and cannot work directly on raw text data because it includes formats that are not acceptable by these algorithms. Methods of feature extraction are used in order to feed the extracted features into machine learning algorithms for learning patterns that can be implemented on new data points [48].

There are different feature extraction methods such as (TF-IDF) and Bag of Words. TF-IDF considers the term frequency and inverse document frequency when weighing each term. However, this thesis adopts TF-IDF model that will be discussed further.

#### **3.5.1 Bag of Words**

Bag of Words (BoW) is one of the best fundamental methods in document classification, used to transform tokens into a set of features. For training the classifier, each word is used as a feature [49] .

(BoW) is called a vector space model. This model works as to convert and represent each document into a vector. This vector refers to the frequency of the words in the specific document. Every row is an observation and every feature is a unique word [49] .

### 3.5.2 Term Frequency-Inverse Document Frequency

TF-IDF is a numerical statistic which shows the significance of the word in the set of documents. TF-IDF is usually used as a weighting factor in text mining and information retrieval. The value of TF-IDF increments relatively to the number of times a word appears in the text, but the repetition of the word in the corpus counteracts it. This can assist in controlling the fact that some words are more well-known than others [50].

TF-IDF is a term derived from the fact that inverted document frequency can be used a numerical statistic that identifies the importance of a word in a set of documents [51]. Mathematically, TF-IDF is the result of the application two scales: TF and IDF. The TF-IDF for each term can thus be calculated using Eq. (1) [51].

$$TD - IDF (t, d) = TF (t, d) \times IDF (t) \quad (1)$$

Where TF is defined as the number of times the term  $t$  occurs in document  $d$ , which is equivalent to the bags of words (BoW) approach. IDF thus refers to the statistical weight used to measure the significance of a term in a set of documents, which can be calculated using Eq. (2) [51].

$$IDF (t) = \log [(n) / (df (t))] \quad (2)$$

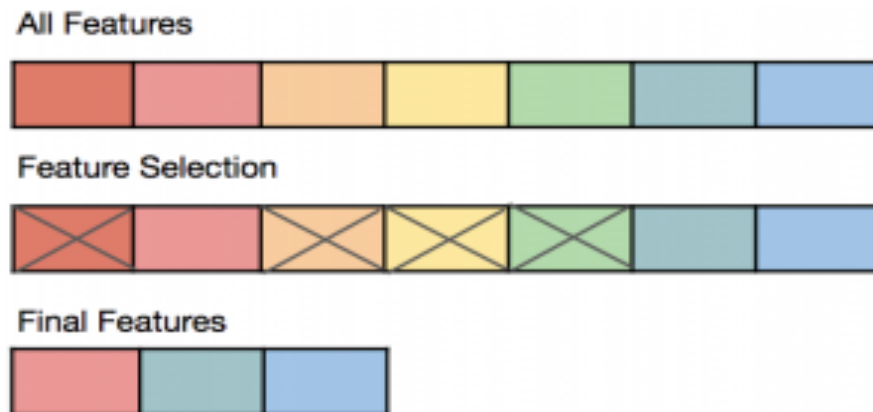
Where the total number of documents in the document set is denoted by  $n$ , and  $df(t)$  indicates the document frequency of  $t$ . Document frequency is thus the number of documents in the set of documents containing the term ( $t$ ).

Once texts are converted using TF-IDF, and given appropriate weights, machine learning algorithms can then be used.

### 3.6 Feature Selection

FS is a process that automatically selects those features in the data that contribute most to the prediction variable or output of interest. FS can thus be used in data pre-processing to achieve efficient data reduction [52]. Sometimes defined as a process of reducing the number of input variables when developing a predictive model. It was intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable [53].

The main goal of using feature selection is an attempt to reduce high dimensionality in features matrix by picking a subset of features as shown in Figure 3.2. High dimensionality can reduce classifier's performance because of the over-fitting and the presence of redundant or useless features. Consequently, the aim of feature selection is to further reduce the dimensionality of the feature set by determining the irrelevant features[54].



*Figure 3.2 feature selection* [55]



There are many feature selection methods such as document frequency thresholding, information gain measure, mutual information measure,  $\chi^2$  statistic measure or Chi-square and term strength measure.

The benefits of feature selection are improving the data quality, increasing the accuracy of the resulting model, improving performance to gain the predictive accuracy, Reducing the dimensionality of the feature space to limit storage requirements and to increase algorithm speed, removing the redundant, irrelevant or noisy data. Improving the effects of data analysis tasks by speeding up the running time of the learning algorithms, reducing feature set to save resources in the next round of data collection or during utilization, and understanding simplify data visualization or gain knowledge about the process of generating the data [56].

After trying several techniques for feature selection, the best feature selection techniques for this work to obtain higher accuracy were identified as the Chi-square test.

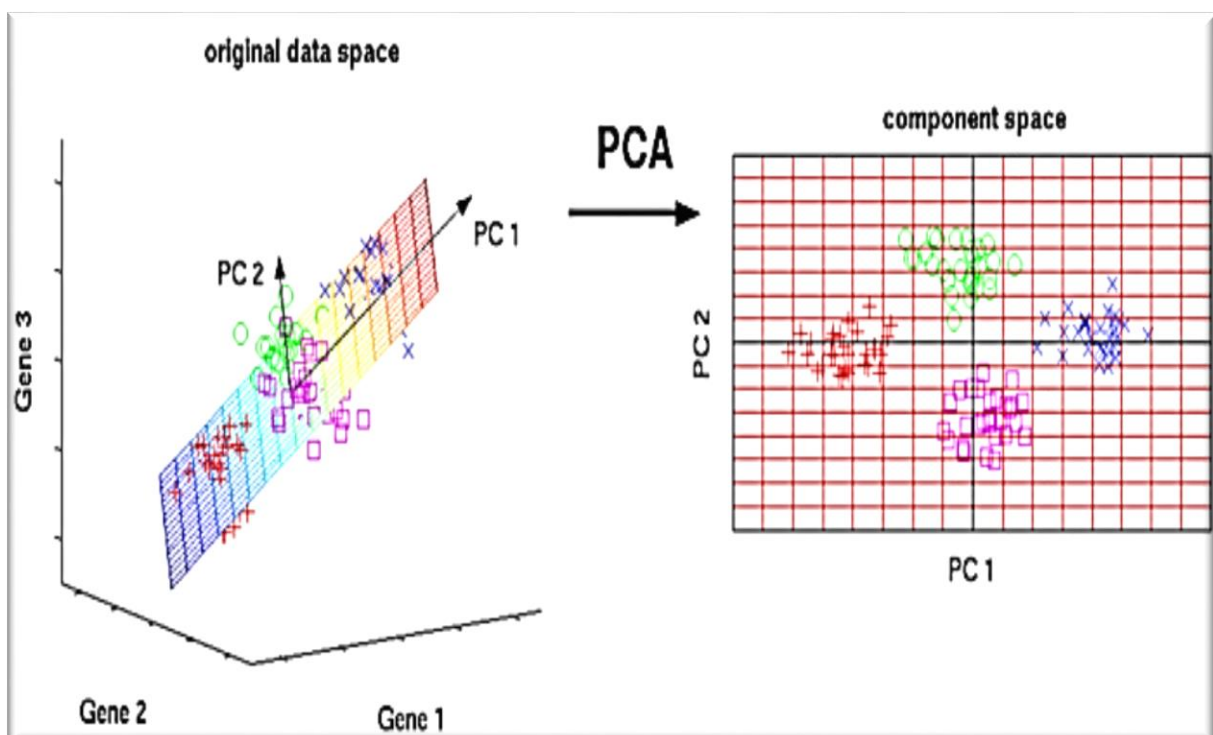
### **3.6.1 Chi-square Test**

The Chi-square test is a common statistical technique used for assessing categorical features in a dataset [57]. The Chi-square value between each feature and the target is calculated, and the desired number of features with the best Chi-square scores thus selected. In order to correctly apply the chi-square test in order to examine the relationship between various features in the dataset and the target variable, the following conditions must be met: the variables must be categorical and sampled independently, and all values should have an expected frequency greater than 5 [58].

### 3.6.2 Principal Component Analysis

(PCA) is a dimensionality reduction technique used to extract features from a dataset by reducing the dimensionality of the dataset based on employing matrix factorisation. This projects the dataset into a lower dimension while attempting to preserve the variance [59].

PCA can be used to reduce the number of features when the dimensionality of a dataset is very high and analysing redundant features is a difficult task. PCA can reduce a dataset with excess features into a dataset with the desired number of features relatively simply, though this does lead to the loss of some variance [60]. Figure 3.3 shows principal component analysis (PCA).



*Figure 3.3 (Principal Component Analysis (PCA)) [61]*

### 3.7 Machine Learning (ML)

ML is the concept of building algorithms that can learn to solve a problem without being directly programmed to solve it. These algorithms employ example data called “Training Data” in order to make data-driven predictions or choices rather than following a firm static program instruction.

Recent years have seen impressive advances in ML, which have raised its capabilities across a suite of applications. Increasing data availability has allowed ML systems to be trained on a big set of subjects [62].

There are two key branches of ML:

- **Supervised Learning:** which requires prior knowledge of the desired outcome. During the learning stage, the ML system must be presented with pairs of example inputs and desired outputs. The learning process will be focused on attempting to guess the output for a specific input, which will then be compared to the actual output.
- **Unsupervised Learning:** which is based on clustering approaches (there is no ground truth information or desired outputs, it is simply a collection of different data. In this case, learning will be focused on finding patterns, which will result in the creation of grouping and various clusters among the data[63]).

There are also other uncommon types of ML like semi-supervised and reinforcement learning. This thesis is based on Supervised ML since it tries to match a MR image to a specific AD diagnosis.

### **3.7.1 Text mining (TM)**

TM is an interdisciplinary field that seeks for extracting significant information from unstructured data. TM is based on data mining (DM), machine learning, information retrieval, statistics, and computational linguistics [64]. Various AI techniques for text mining are used to automatically process data and generate useful or valuable insights, enabling users to make decisions based on the information extracted from such data [65].

Research in text mining field is still active because there is a massive number of information such as Web pages, books, technical papers, digital libraries, blogs, email messages, and news articles. Consequently, a significant goal of text mining is to be able to obtain high-quality information from text.

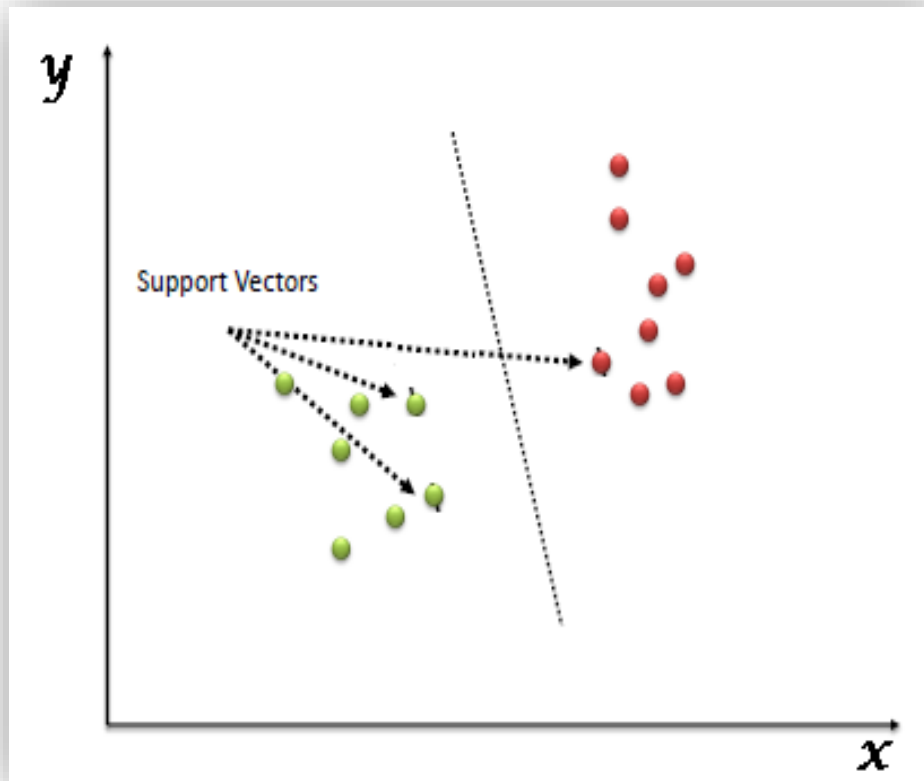
TM is similar to data mining, but the only difference is that text mining is used with semi-structured or unstructured datasets such as emails, HTML files, texts, etc., while data mining tools are used to manipulate the structured data from databases [66]. Text mining deals with natural language text which is in semi-structured and unstructured format. Many text mining methods can be used in order to extract knowledge such as classification, clustering, and text summarization (information extraction). TM identifies facts, assertions, and relationships buried in blocks of textual big data that, without excavation, would never be discovered and would remain buried indefinitely. Excavation can, however, extract the useful information, and then transform it into a structured model that can be analysed further or even presented directly without analysis. In terms of data classification, different techniques, such as decision tree and logistic regression methods, have been proposed [67].

Although many other data mining techniques have been utilised in the literature, only the methods most relevant to this thesis are presented here [64].

### 3.7.1.1 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm in which each data element is plotted as a point in an  $n$ -dimensional space representing  $n$  available features, with the value of each feature being a given coordinate value. The classification process is achieved by finding the hyper-level that most clearly distinguishes between the two categories, as shown in Figure 3.4 [68].

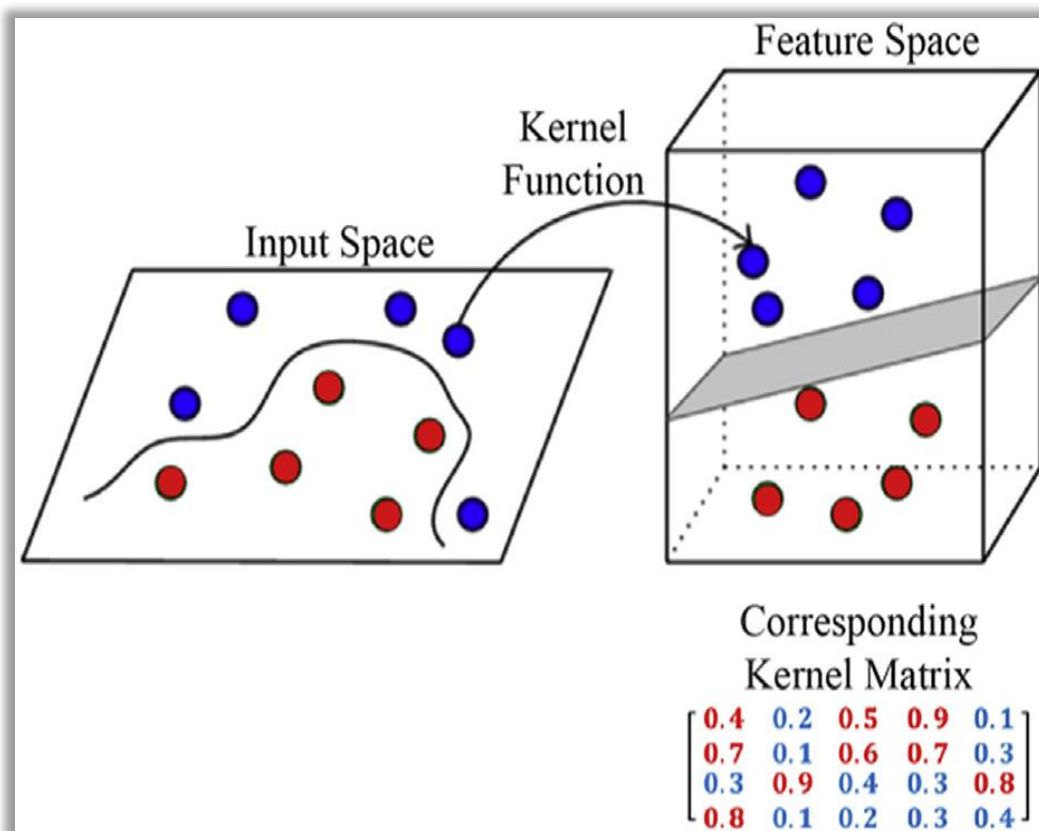
SVM was invented by Vapnik (1995) based on the theory of statistical learning, it was originally designed for pattern recognition and multidimensional regression estimation, which can be used to solve linear or nonlinear problems [69].



*Figure 3.4 Illustration of a hyperplane [70]*

SVM is only effective in high-dimensional spaces, though it remains effective in cases where the number of samples is less than the number of dimensions. SVM is efficient in terms of memory use because it uses a subset of training points in the decision function (support vectors), and it is versatile in that it is possible to define a custom kernel as well as various kernel functions for the decision function, as shown in Figure 3.5.

It is ineffective if the number of features is significantly greater than the number of samples, and it cannot provide probabilistic estimates directly [71]. In the context of NLP, there are hundreds or thousands of features in most training data; thus, using a non-linear form to map the data in the input to a large space may offer the best performance in these cases.

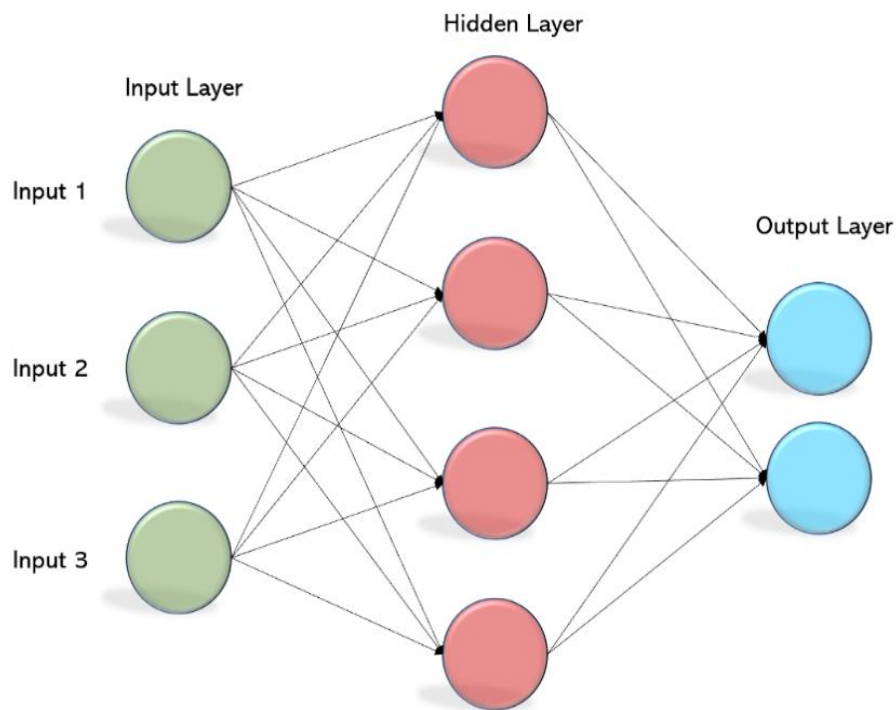


**Figure 3.5 Support Vector Machine [70]**

### 3.7.1.2 Multiple Layer Perceptron

Multiple Layer Perceptron (MLP) is an artificial neural network (ANN) with one or more hidden layers where the nodes in each layer are made up of nonlinearly activated neurons. MLP thus consists of one or more input layers, one or more hidden layers, and output layers where outputs from nodes are interconnected in a feed-forward direction [72].

MLP uses backpropagation (BP) technology for training that consists of redirection and back feed phases. Figure 3.6 shows that the input is fed in the first stage to the nodes of the first hidden layer to perform activities related to the activation function passing from the input layer to the output layer, while in the second stage, the error between the desired and actual value is used to adjust the learning weights based on spreading the input layer from the output layer [73].



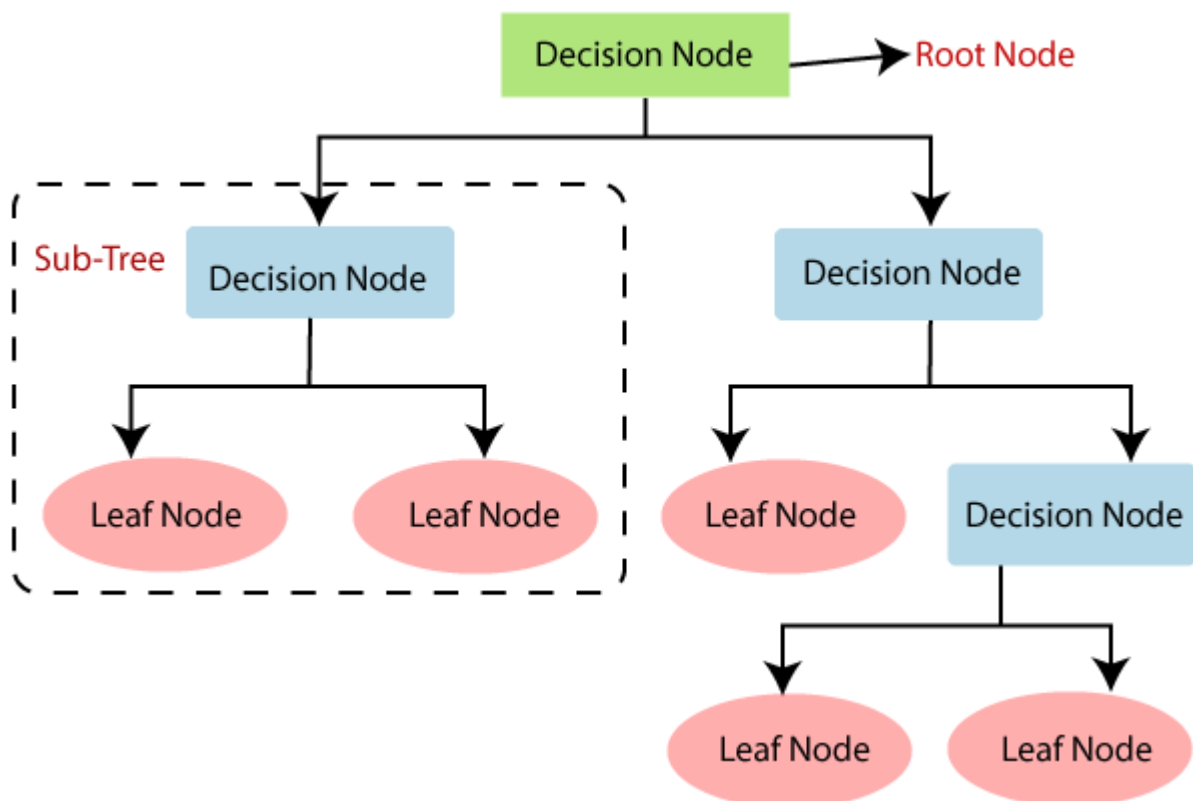
**Figure 3.6 Multiple Layer Perceptron (MLP) [74]**

### 3.7.1.3 Decision Tree

Decision tree (DT) technique is an effective way of generating models in the form of a tree structure [75]. DT method breaks down a dataset into smaller and smaller subsets. At the same time, however, an associated decision tree is incrementally developed. The result in this technique is a tree with decision and leaf nodes. DTs can handle both categorical and numerical data. Figure 3.7 illustrates the main notion behind decision tree algorithm. The rule for decision tree is shown in Equation (3) [76].

$$Entropy(t) = - \sum^{c-1} p(i|t) \log_2 p(i|t) \quad (3)$$

Where  $c$  is the number of classes,  $p(i | t)$  denotes the fraction of records belonging to class  $i$  at a given node  $t$



*Figure 3.7 Decision Tree* [77]



### **3.7.2 Deep learning**

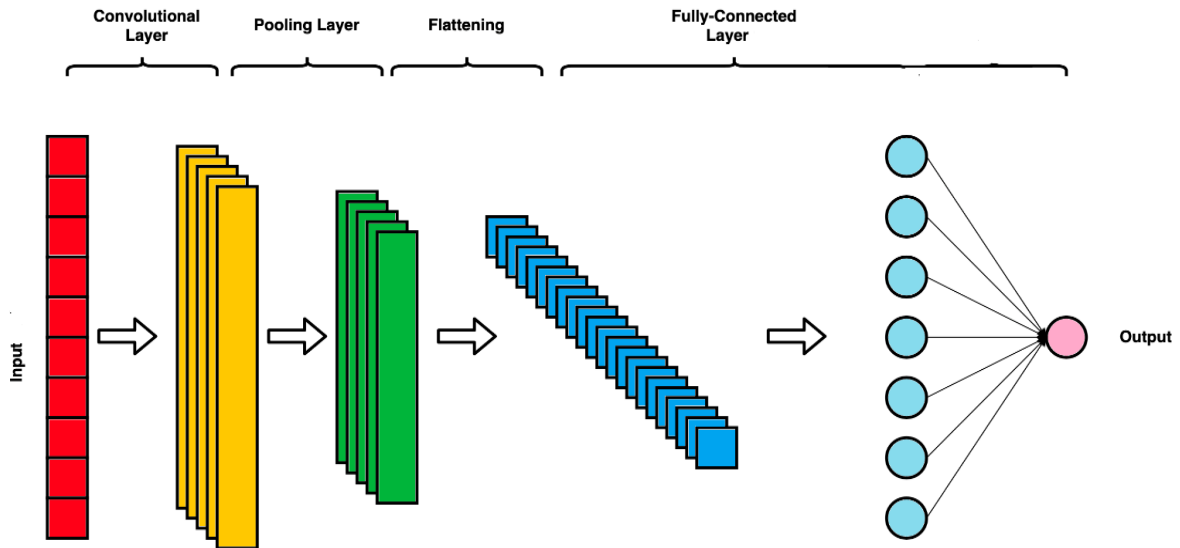
(DL) is a set of algorithms and techniques hierarchical machine learning inspired by how the human brain works, called neural networks. DL architectures offer a big benefit for text classification because they work at very high accuracy with lower-level of computation [78].

The concept of ‘deep learning’ has recently gained a lot of attention. It refers to unsupervised learning algorithms which automatically discover data without the need for supplying specific domain knowledge [79]. This approach usually has higher performance rates than supervised and informed methods when processing large unstructured corpora. However, the ability of these algorithms has to be evaluated to measure their error rate.

Deep learning-based models have surpassed classical machine learning-based approaches in various text classification tasks, including sentiment analysis, news categorization, question answering, and natural language inference[80].

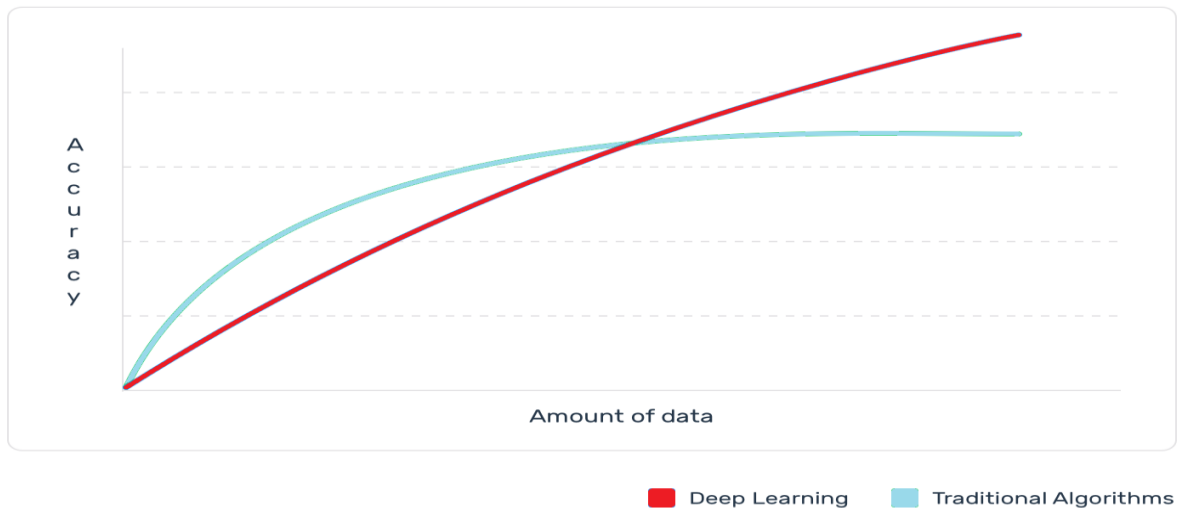
A large amount of biomedical information in databases such as EMRs is a valuable source for information extraction [81]. Information extraction and, in particular, natural language processing methods are required to annotate and process biomedical literature [82].

There are two main architectures for deep learning text classification, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Figure 3.8 shows simple of deep learning layers.



**Figure 3.8** simple of convolution neural networks layers

DL algorithms require big training data (at least millions of tagged) compared with traditional machine learning algorithms. DL algorithms do not have a threshold for learning from training data, like traditional machine learning algorithms, such as Support Vector Machine and DT classifiers continue to get better the more data you feed them with as shown in Figure 3.9.



**Figure (3:9)** The Relationship Between Accuracy and Amount of Data [83]

### 3.7.3 One-hot Full Embedding

This is the most straightforward and commonly used approach to embed categorical features, which assigns each feature value a unique  $d$ -dimensional embedding in an embedding table as shown in Figure 3.10-A. Specifically, the encoding function  $E$  maps a feature value into a unique one-hot vector. the embedding lookup process can be viewed as a 1-layer neural network (without bias terms) based on the one-hot encoding [84].

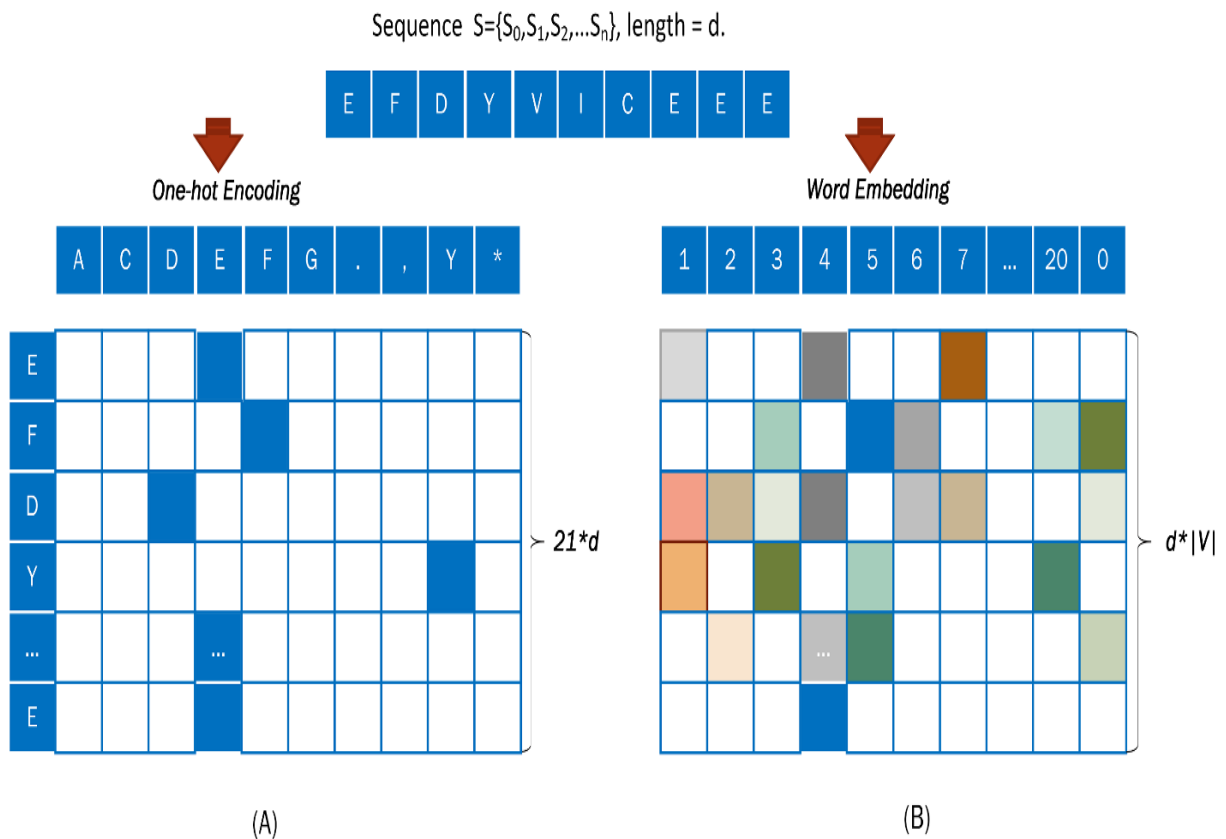
### 3.7.4 Word Embedding

Word embedding is a feature learning technique that is foundational to natural language processing a great asset for a large variety of natural language processing (NLP) tasks, it is a means of turning texts into numbers. We do this because machine learning algorithms can only understand numbers, not plain texts. Word embedding means representing the words of the text in an  $N$ -dimensional vector space, as shown in Figure 3.10-B thereby enabling the capture of semantics, the semantic similarity between words, and syntactic information for words [85].

Aim of Word embedding is mapping words from the vocabulary into vectors of real numbers in a low-dimensional space, by leveraging large corpora of unlabeled text such continuous space representations can be computed for capturing both semantic and syntactic information about words [31] , [78]. Embedding then make grouping of the commonly co-occurring items together in the representation space.

Embedding is a method that requires large amounts of data and a long training time. The result is a dense vector with an arbitrary number of dimensions. Therefore, if there is enough training data, enough training time, and the ability to apply the more complex training algorithm (e.g., word2vec or GloVe), go with Embedding.

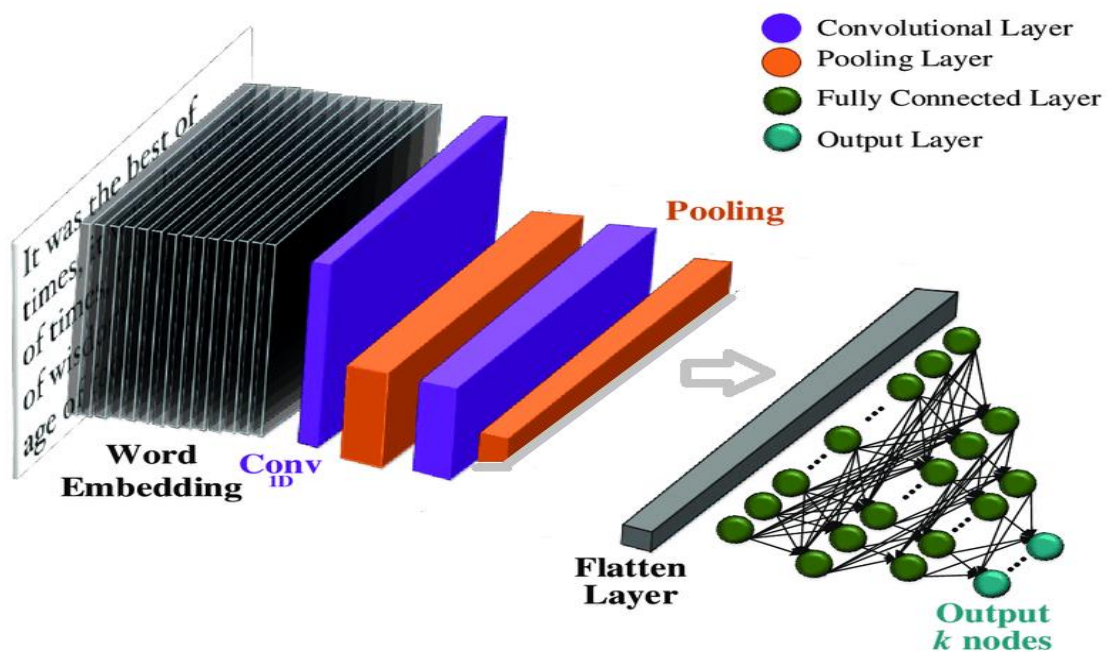
Embedding is used to represent discrete variables as continuous vectors. It produces a dense vector with a fixed, arbitrary number of dimensions. Word embedding is one of the most popular representations of document vocabulary. The word embedding representation can reveal many hidden relationships between phrases [86].



**Figure 3.10 Coding diagram of (A) One-hot encoding and (B) Word embedding encoding [86]**

### 3.7.5 Convolutional Neural Networks (CNNs)

ConvNets are designed to process data that come in the form of multiple arrays. There are four key ideas behind ConvNets that take advantage of the properties of natural signals: local connections, shared weights, pooling and the use of many layers [87]. Figure (3.11) shows simple of convolution neural networks for text document.



*Figure 3.11 simple of convolution neural networks for text document [88]*

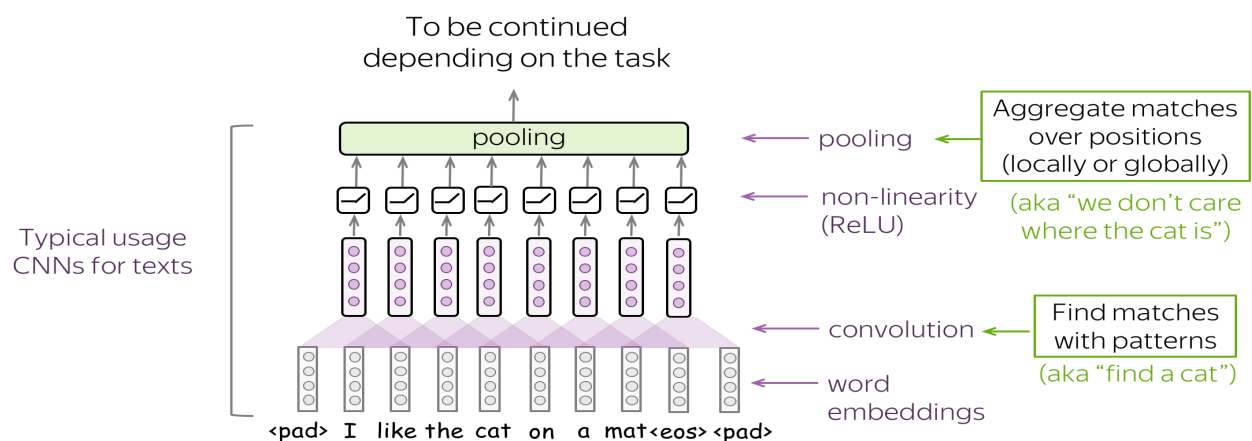
(CNN) was the best types of the deep neural networks and most important and useful, typically used in classification of text and object segmentation in images. CNN consists of three main layers: convolution layer, pooling layer, and fully connected layer. Each of these layers does certain spatial operations. In convolution layers, CNN uses different kernels for convolving the input image for creating the feature maps. The pooling layer is usually inserted after a convolution layer. The application of this layer is reducing the size of feature maps and network parameters.

After the pooling layer, there is a flatten layer followed by some fully connected layers. In the flatten layer, 2D feature maps produced in the previous layer are converted into 1D feature maps to be suitable for the following fully connected layers. The flattened vector can be used later for the classification of the text.

One advantage of CNNs is the weight sharing mechanism due to the use of the kernels, which results in a smaller number of parameters than a similar fully-connected neural network, making CNNs very easy to train.

### 3.7.5.1 convolution layer

Convolution finds matches with patterns, texts have only one dimension, Therefore, a convolution here is one-dimensional. A typical convolutional model for texts is shown on the Figure 3.12. Usually, a convolutional layer is applied to word embedding, which is followed by a non-linearity (usually ReLU) and a pooling operation. These are the main building blocks of convolutional models: for specific tasks, these blocks are standard.

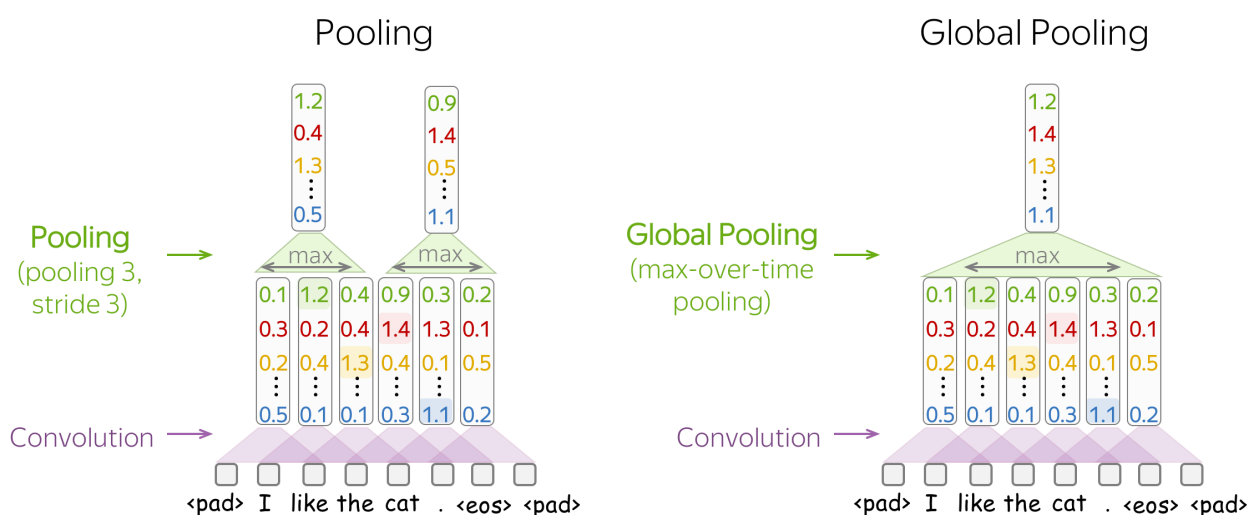


**Figure 3.12 typical convolutional model for texts [89]**

### 3.7.5.2 pooling layer

After a convolution extracted features from each window, a pooling layer summarizes the features in some region. Pooling layers are used to reduce the input dimension, and, therefore, to reduce the number of parameters used by the network [90].

For texts, global pooling is often used to get a single vector representing the whole text; such global pooling is called max-over-time pooling, where the "time" axis goes from the first input token to the last. As shown in Figure (3.13).

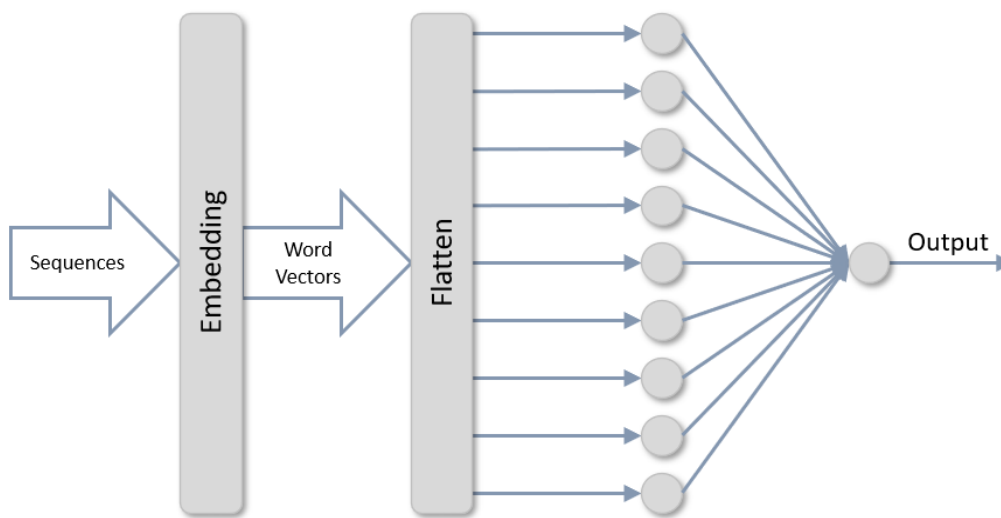


**Figure 3.13 simple of pooling layer [89]**

Max-pooling induces a thresholding behavior. The values below a given threshold are ignored irrelevant to make a prediction. Sometimes 40% of the pooled n grams on average can be dropped with no loss of performance.

### 3.7.5.3 flatten layer and fully connected layer

Flattening is converting the data into a 1-dimensional array for inputting it to the next layer as shown in Figure 3.14. flatten the output of the convolutional layers to create a single long feature vector and it is connected to the final classification model, which is called a fully-connected layer. In other words, put all the data in one line and make connections with the final layer [91].

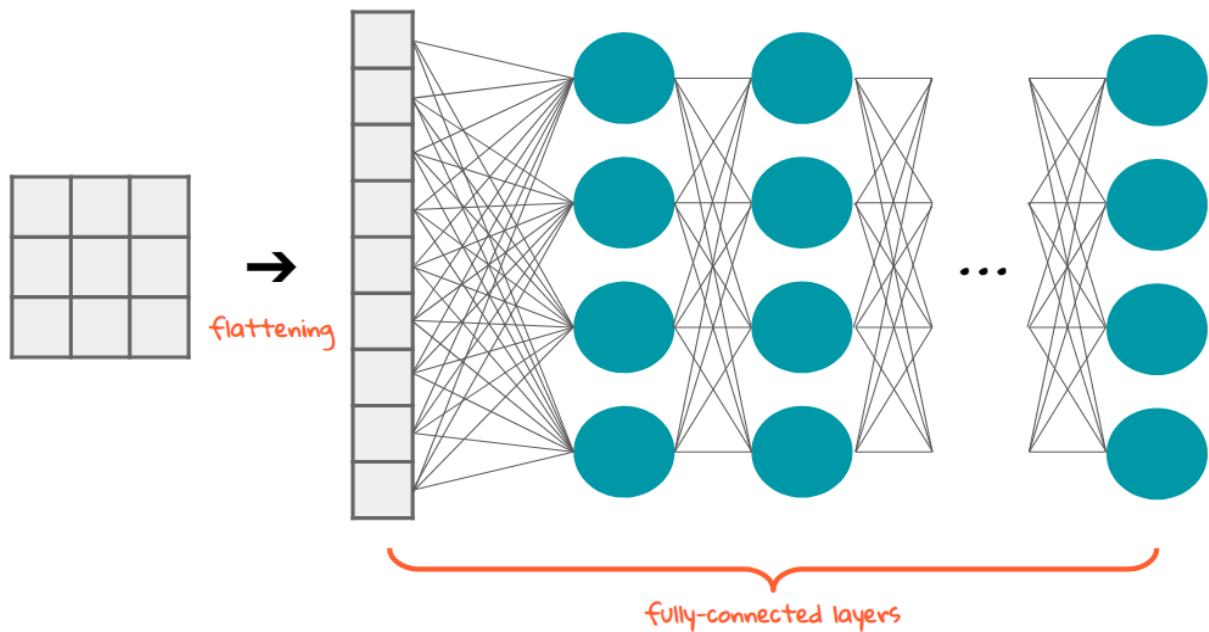


*Figure 3.14 Embedding layer and flatten layer [92]*

### 3.7.5.4 Fully-Connected Layer

The fully-connected layer also known as the dense layer. all the resulting features that selected from the max-pooling layer are combining. the max-pooling layer selects the k-most feature from each convolutional kernel. The fully-connected layer can combine most of the useful assemble and then construct a hierarchical representation for the final stage, the output layer [93]. Figure (3.15) shows the flatten layer and the fully connected layer.





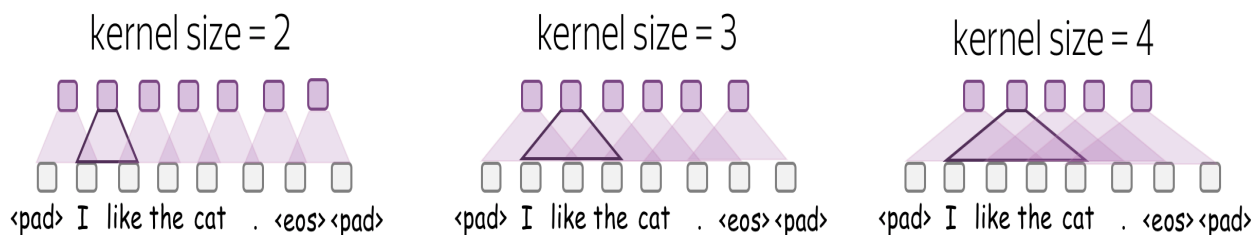
*Figure 3.15 flatten layer and fully connected layer [94]*

### 3.7.5.5 Dense layer

Dense layer is a layer that is deeply connected with its preceding layer which mean the neurons of the layer are connected to every neuron of its preceding layer. This layer is the most commonly used layer in artificial neural networks. The dense layer's neuron in a model receives output from every neuron of its preceding layer, where neurons of the dense layer perform matrix-vector multiplication. Matrix vector multiplication is a procedure where the row vector of the output from the preceding layers is equal to the column vector of the dense layer. The general rule of matrix-vector multiplication is that the row vector must have as many columns as the column vector.

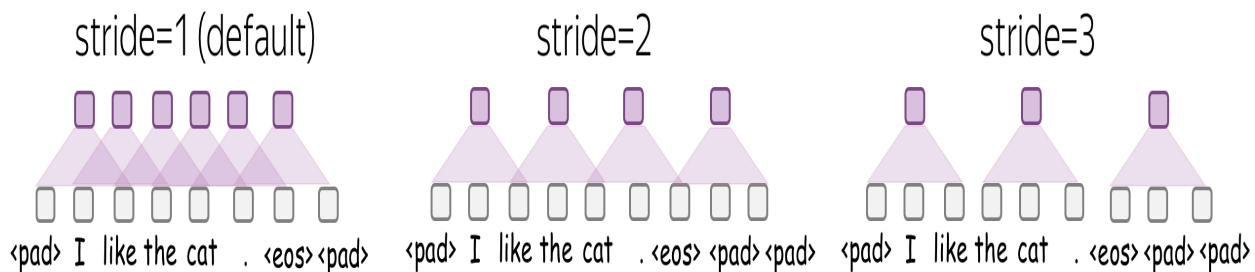
### 3.7.6 Parameters

**Kernel size** is the number of input elements (tokens) a convolution looks at each step. For text, typical values are 2-5. As shown in Figure 3.16.



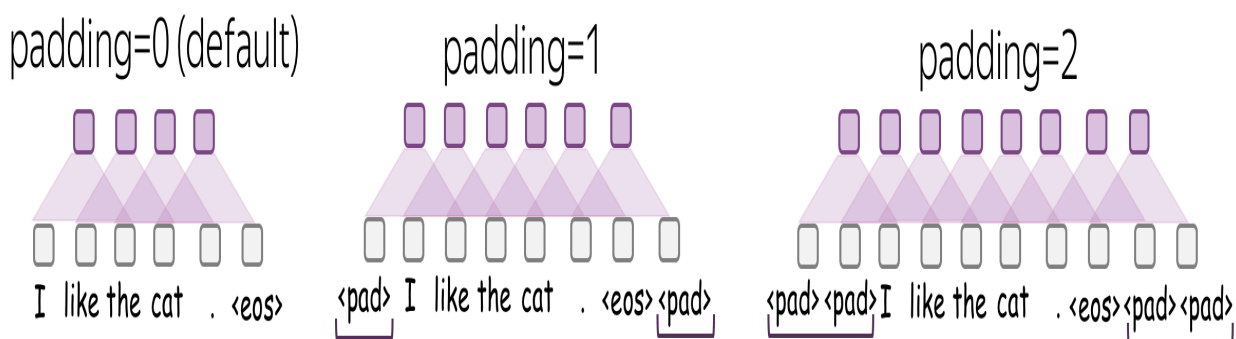
**Figure 3.16 Kernel** [89]

**Stride** tells how much to move filter at each step. For example, stride equal to 1 means that we move the filter by 1 input element (pixel for images, token for texts) at each step. As shown in Figure 3.17.



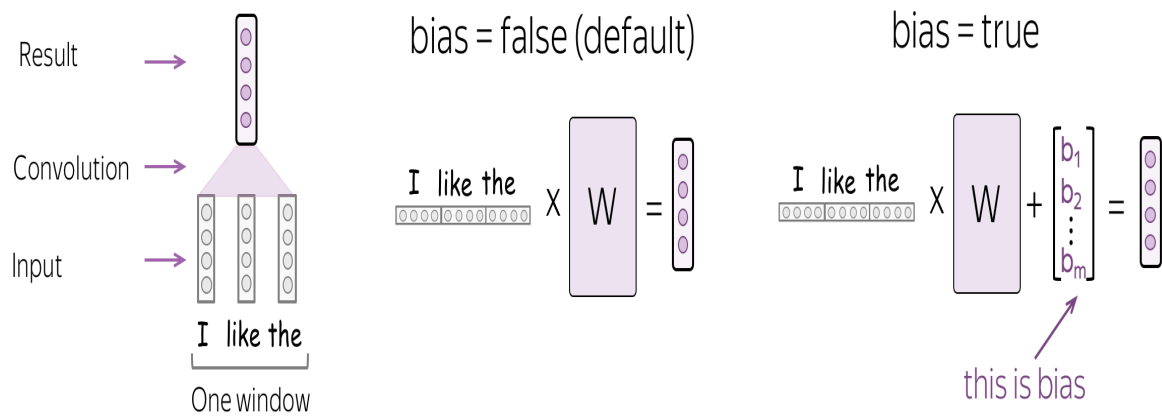
**Figure 3.17 Stride** [89]

**Padding** adds zero vectors to both sides of an input. As shown in Figure 3.18.



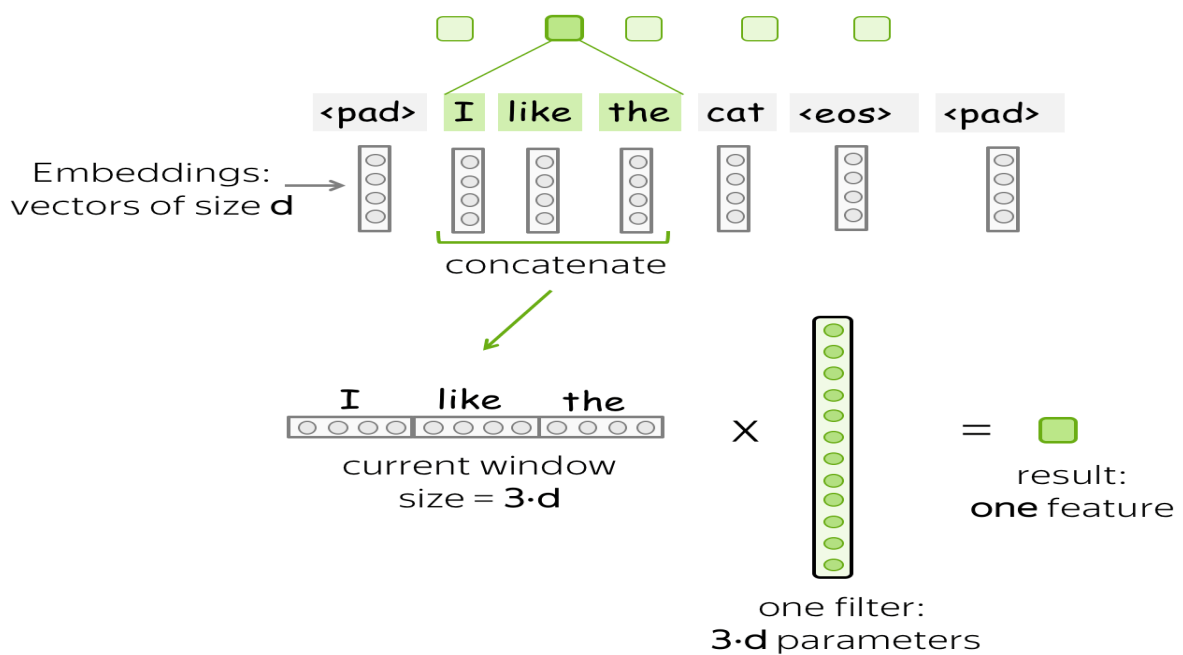
**Figure 3.18 Padding** [89]

**Bias** The bias term in the linear operation in convolution, By default, there's no bias - only multiplication by a matrix. As shown in Figure 3.19.



**Figure 3.19 Bias** [89]

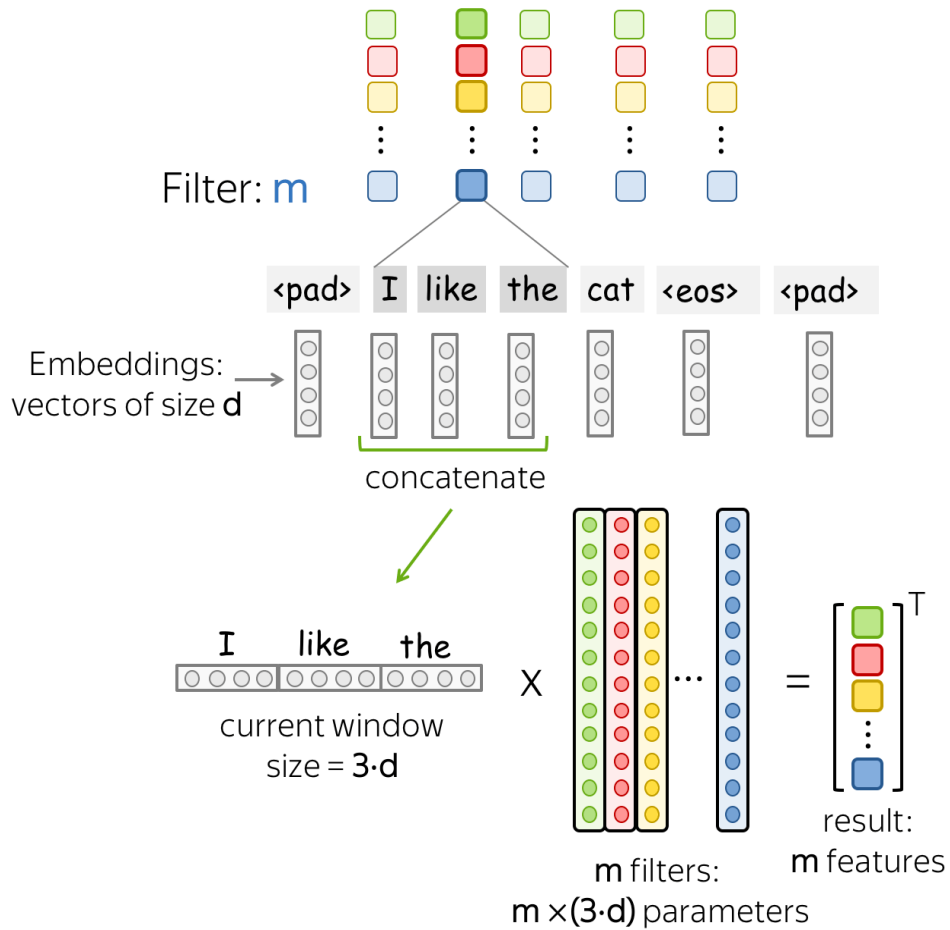
**Intuition:** Intuitively, each filter in a convolution extracts a feature (One filter - one feature extractor), A filter takes vector representations in a current window and transforms them linearly into a single feature.



**Figure 3.20 Intuition** [89]

## Filters.

One filter extracts a single feature. Thus, to extract many features, must take several filters. Each filter reads an input text and extracts a different feature. The number of filters is the number of outputs features you want to get. Figure 3.21 shows the filters with  $K=3$  [95] .



*Figure 3.21 filters* [89]

### 3.8 Performance Metric for Classification Algorithms

Several different metrics can be used to evaluate the efficiency of particular classification algorithms based on assessing accuracy, f1-score, precision, and recall. The calculation of these measures is based on the computing confusion matrix, which is a matrix that summarizes the number of examples correctly or incorrectly predicted by a classification model, as discussed in more detail in table 3.1 below.

*Table 3.1 Confusion Matrix*

		Predicted Class	
		Positive +	Negative -
Actual Class	Positive +	$f_{++}$ (TP)	$f_{+-}$ (FN)
	Negative -	$f_{-+}$ (FP)	$f_{--}$ (TN)

Were,

1. **True positive (TP)**: denotes to the positive examples that are properly classified.
2. **False negative (FN)**: denotes to the positive examples that are incorrectly classified.
3. **False positive (FP)**: denotes to the negative examples that are incorrectly predicted and classified.
4. **True negative (TN)**: denotes to the negative instances that are properly predicted by the classification model.

In this thesis,

1. True positive (TP): Number of positive cases correctly diagnosed.
2. True negative (TN): Number of negative cases correctly diagnosed.
3. False negative (FN): Number of positive cases incorrectly diagnosed as negative
4. False positive (FP): Number of negative cases incorrectly diagnosed as positive.

The confusion matrix values can be used to evaluate the model performance based on test dataset to obtain the following metrics:

### 3.8.1 Accuracy

Accuracy refers to the overall accuracy of the model in its entirety, which can be calculated by calculating the total number of correct classifications and dividing this by the total number of classifications made, as shown in Eq. (3): In this case, accuracy represents the number of correctly classified recordings divided by the total number of recordings.

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

### 3.8.2 Precision

Precision is a measure of accuracy where a specific class has been predicted. Precision is thus related to the probability that a retrieved document is relevant. From the confusion matrix, this can be calculated by using Eq. (4):

$$PRECISION = \frac{TP}{TP+FP} \quad (4)$$

where TP and FP are the numbers of true positive and false positive predictions for the considered class, respectively. Precision is 1 when FP is 0, as this indicates there are no spurious results.

### 3.8.3 Recall

Recall is the probability that a relevant document will be retrieved in a search. Recall is also referred to as the true positive rate or sensitivity, and it is given by Eq. (5):

$$RECALL = \frac{TP}{TP+FN} \quad (5)$$

Recall becomes 1 when FN is 0, indicating that 100% of the TP have been discovered.

### 3.8.4 . F-measure

The F-measure is the harmonic mean of precision and recall, calculated by using the formula in Eq. (6):

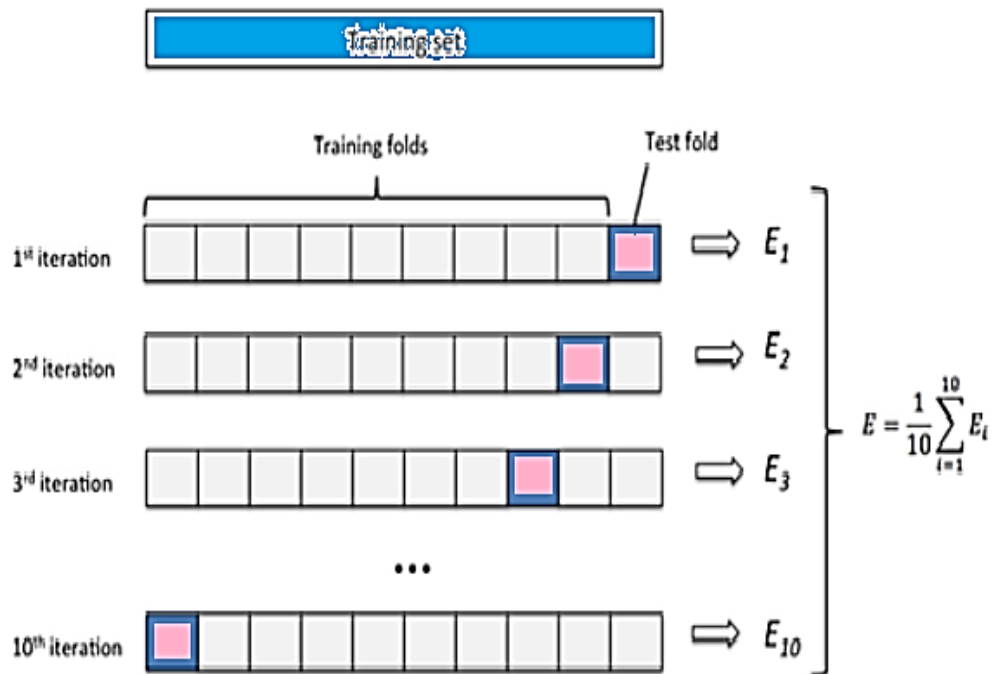
$$F1-score = \frac{2*TP}{2*TP+FN+FP}$$

*Or*  $F1-score = 2 * \frac{Precision \times Recall}{Precision + Recall} \dots\dots\dots(6)$

The behaviour of this performance measure is the function of the decision threshold for classification. When decision threshold increases, the recall will increase, and precision will decrease.

### 3.8.5 Cross Validation

Cross-validation is a model validation technique used to assess how results from a dataset generalize to independent data sets. This is performed on the samples of data, which are partitioned usually into a training and validation sets. There are many different methods to perform cross-validation such as leaving one out and k-fold [96]. The best type of cross-validation when using 10-fold cross-validation. 10-fold cross-validation splits our dataset into 10 random sub-sets where 1 of which will be used for training, and 9 of which will be used for testing. This process is repeated 10 times until all permutations are used for training and testing as shown in Figure 3.22, k-fold cross validation where  $k = 10$ , [96].



*Figure 3.22. 10-Fold Cross Validation [97]*



# CHAPTER 4

## METHODOLOGY

### 4.1 Introduction

Large datasets of detailed information about patients, including disease status, medication history, and side effects, treatment outcomes, and laboratory test results, are collected in an electronic format. This is generally referred to as an EMR, and the data serves as a valuable resource for further analysis, diagnosis, and treatment. A large amount of detailed patient' information in these medical texts produces a big challenge in terms of processing this data efficiently.

ML algorithms, AI and NLP tools have the potential effect of simplifying unstructured data, which could positively affect medical report analysis. NLP has recently made significant advances, outperforming classical statistical and rule-based systems on a variety of tasks. In this thesis, an automatic system was produced to classify specialist consultant interactions based on patients' medical reports.

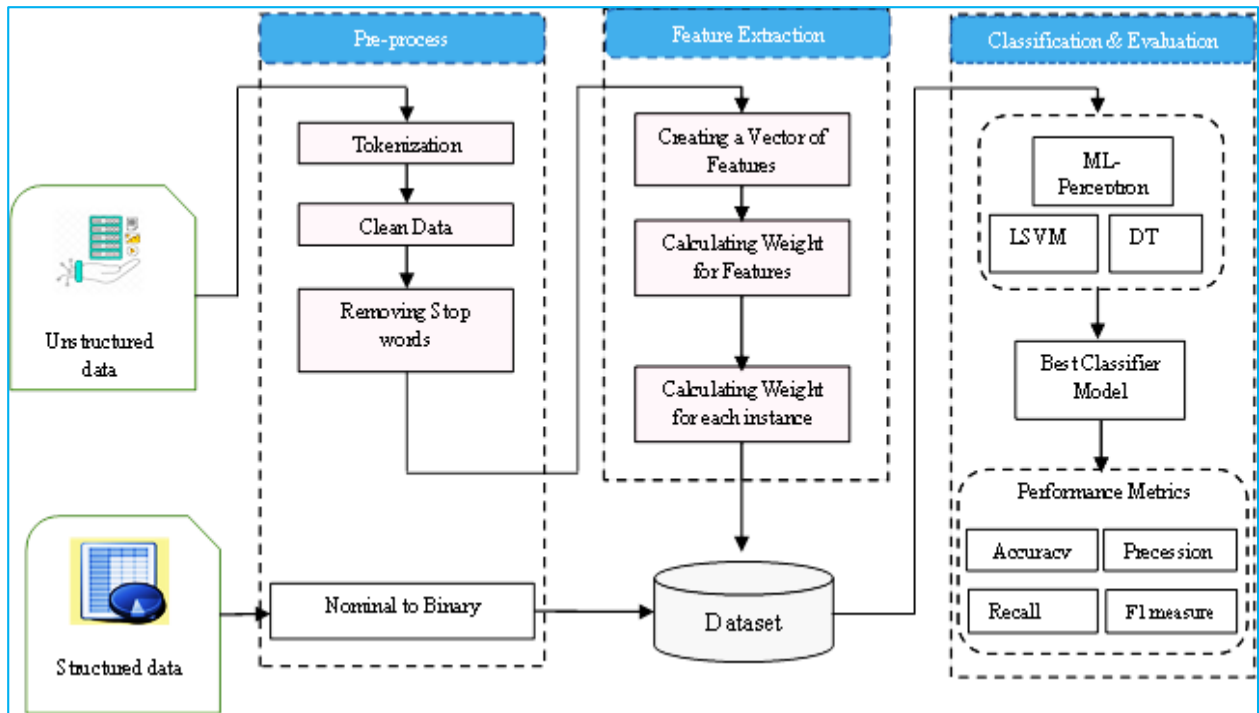
The natural language segment in all models include manual spell checking, tokenization, and data cleaning, as well as removing all English and domain stop words to help reduce the number of features in the report and to eliminate various noise steps.

## **4.2 The Proposed System Architecture**

In this thesis, three main proposed models of diseases diagnosing and medical report categorization conducting on semi structured and unstructured data respectively. First model was diagnosed asthma disease from semi structured data by using classical methods of classifier. the second model was multi classification of medical text report by using feature extraction and feature selection and classical method for classifier, the last model used deep learning algorithms to classify medical text reports into multi classes.

### **4.2.1 Asthma diagnosis model**

The proposed system architecture includes four stages. Each stage consists of sub-steps in order to achieve the research objectives and meet its main aim. the first stage is data collection, the second stage is data pre-processing, the third stage is a feature extraction and the last stage is classification and evaluation, as shown in Figure 4.1. The first stage is data collection, two types of data are used in this step, Grenada dataset and local dataset, for the data collected locally, the data must be saved in an excel file in CSV format so that the model can deal with it. The second stage is data pre-processing which encompasses many processes to prepare the proposed system input. The third stage is feature extraction which involves the construction of vocabulary from all words that are extracted from medical text reports. This also involves creating a vector of features and calculating the weight of each feature by using TF-IDF. Classification methods that are applied are represented in the final stage. This encompasses the application of Decision Tree (DT), Random Forest (RF), and Linear Support Vector classifiers (LSVC).



*Figure 4.1 Asthma diagnosis model.*

#### 4.2.2 Proposed System for classical classifier model

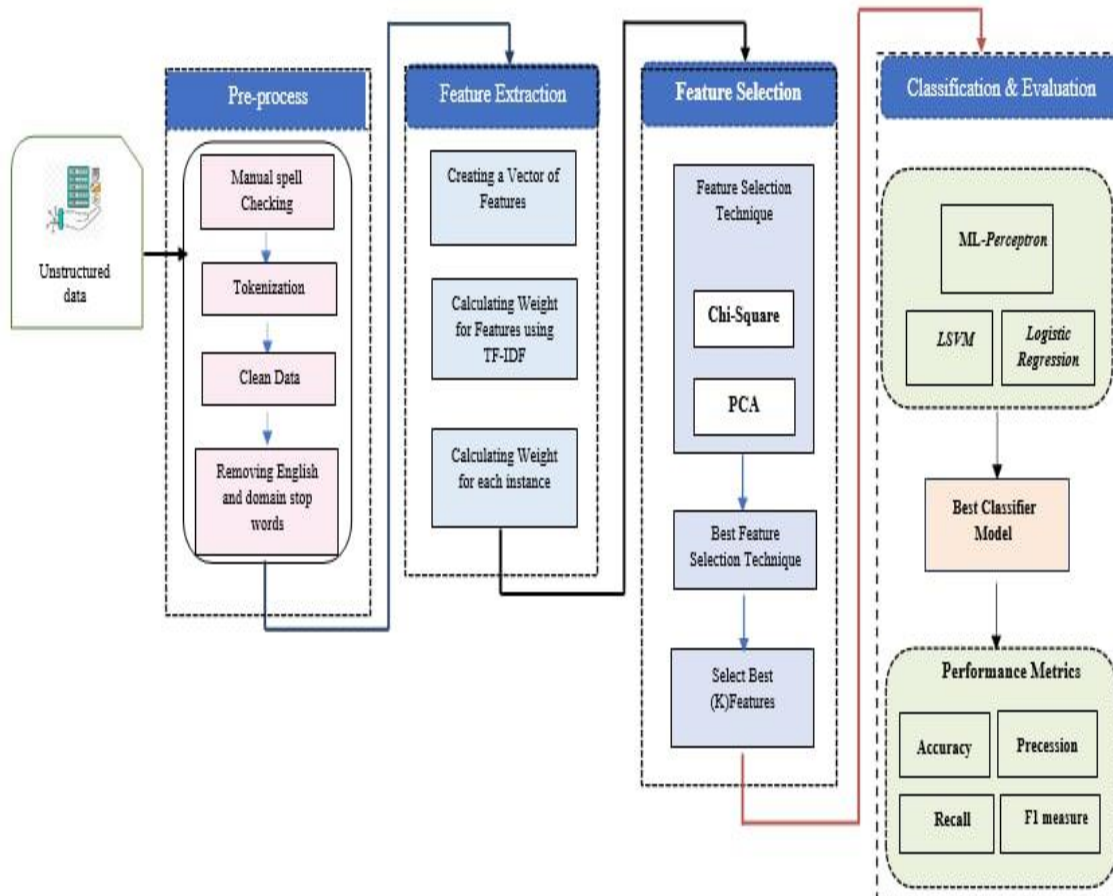
The first proposed system architecture includes five stages. Each stage consists of four sub-steps in order to achieve the research objectives and meet its main aim.

These stages are data pre-processing, feature extraction techniques and feature selection techniques, classification, and evaluation, as shown in Figure 4.2.

The first stage is data pre-processing which encompasses many processes to prepare the proposed system input.

The second stage is feature extraction which involves the construction of vocabulary from all words that are extracted from the text report. and it involves creating a vector of features and calculating the weight of each feature by using TF-IDF.

The third stage comprises feature selection which helps reduce the number of features used in the classification stage. Classification methods that are applied are represented in the fourth stage. This encompasses the application of DT, RF and LSVC.



**Figure 4.2 Proposed System for classical classifier model**

### 4.2.3 Deep learning Classifier Model

In this model, two stages have been applied, the first stage is preprocessing stage, the same as preprocessing in the first model with addition of the Encoder stage using (one-hot encoding), in this stage the categorical variables in text reports are converted into a numerical form to enable algorithms to deal with them. The second stage was CNN deep learning. It contains five Layers, the first layer is word embedding, the learning process of Word embedding methods was joined with the neural network model on task, and the second layer was CNN. Then we used the pooling layer, then flatten layer, And the last layer was a fully connected layer. Figure 4.3 shows the proposed System of Deep Learning Classifier.

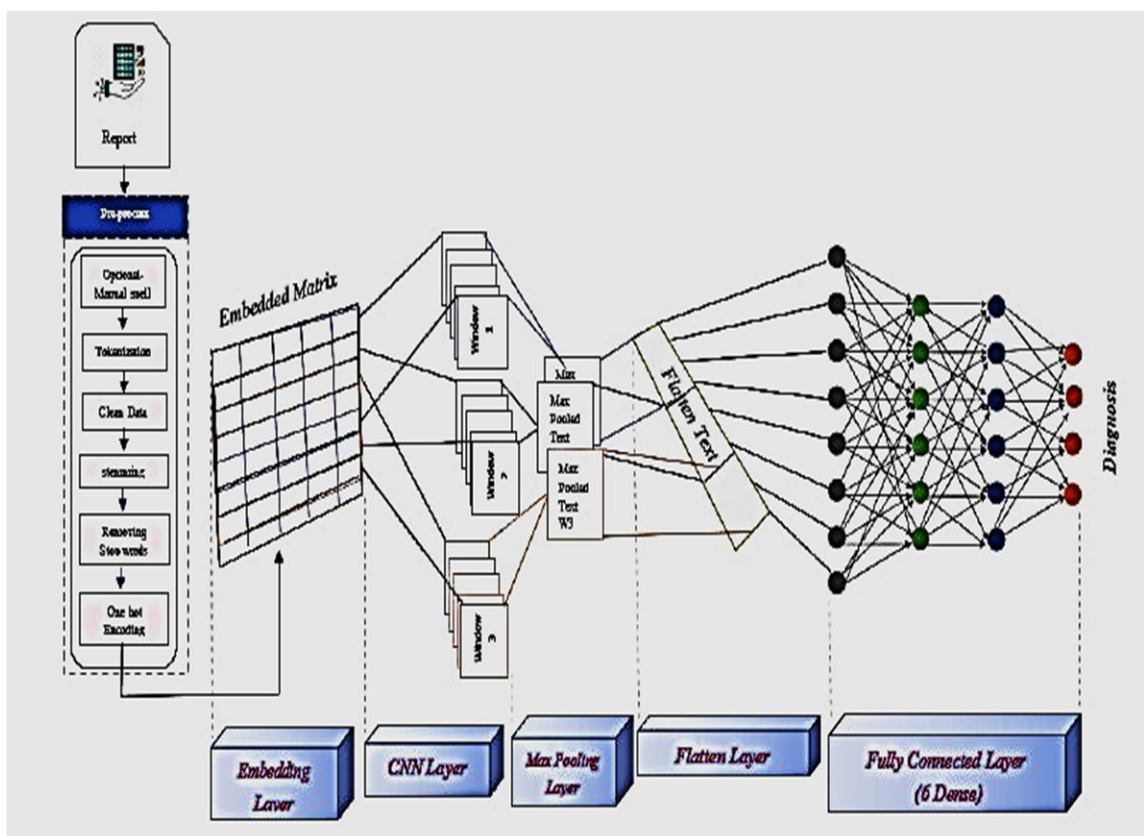


Figure 4.3 Propose System of Deep Learning Classifier

### 4.3 Datasets

The first data was collected for patients with Asthma and Allergies in Childhood from study done in the University of Grenada from an asthma survey across Grenada in the West Indies. Surveys include 1,165 cases. for children's ages between 6 and 8 years old, the data collected included various aspects of information about asthmatic patients including relatives with asthma and the treatment suggested by the relevant doctors.

Local database was established from 201 survey cases were gathered for about one year. These data were collected by means of a survey conducted among a group of consultants specializing in chest and respiratory diseases at Imam Hussein medical city which is a teaching hospital in the holy Karbala Governorate. The survey sheets were thus filled by pulmonologists then save the results in CSV format.

The second data used in the second and third model is a medical record database contains self-information, symptoms and medication notes details of a large number of patients collected during ten years from American EMRs downloaded in CSV format. Two columns were chosen from dataset in this thesis first column is medical report and the second column is medical speciality as shown in table 4.1.

**Table 4.1 Sample of the EMRs Dataset**

NO.	EMR Sample
1	s:33 yr old female crystallographer presents today for routine exam. patient reports no acute problems. Patient reports that she never drinks alcohol. she denies smoking. o:Height 160 cm, Weight 53.8 kg, Temperature 37.3 C, Pulse 76, SystolicBP 146, DiastolicBP 93, Respiration 15, HPV I/H Risk DNA Probe negative for HPV 16 & 18, Visual Acuity Study right eye 20/20, left eye 20/20 a:normal exam. no current issues. problem status: Hypertension, being managed. administered immunization: FLUARIX p:Call office if any reaction from immunization. F/Up in one (1) year for annual check-up or sooner for new symptoms/problems as they arise.
2	s:a 32 year old f presents with critical dyspnea, critical shortness of breath and critical cough. Patient reports that she never drinks alcohol. patient has a one pack per day habit. o:Height 173 cm, Weight 91.1 kg, Temperature 37 C, Pulse 92, SystolicBP 142, DiastolicBP 91, Respiration 14, FEV1 FEV1=35 %, FEV1/FVC FEV1/FVC=60 %, Arterial Blood Gas PaCO2=44 mmHg,PaO2=58 mmHg, plum: accessory muscle use,
3	s:a white female aged 32 Ys presents with 8 months history of mild spells of vertigo. pt also reports increased frequency of mild ringing in the ears, mild headaches particularly at the back of the head and in the morning. o:Height 173 cm, Weight 93.2 kg, Temperature 37.1 C, Pulse 93, SystolicBP 147, DiastolicBP 94, Respiration 16, Heart = 2/6 systolic murmur at base of heart, Chest = clear to auscultation B/L, no rales or wheezing, Extremities = no edema or clubbing, Heart = normal S1, S2, RRR a:Hypertension p:performed E/M Level 2 (established patient) - Completed, and prescribed Hydrochlorothiazide - 50 mg po qd, and ordered Basic Metabolic

This data was available online as a public, free-to-use dataset. It contains data about 3318 medical encounters. Table 4.2 shows the distribution of dataset.

**Table 4.2 the distribution of dataset.**

<i>Dataset</i>	<i>Quantity</i>
<i>Number of sentences</i>	<i>15153</i>
<i>Number of unique words</i>	<i>1935</i>
<i>Number of training</i>	<i>2322</i>
<i>Number of tested</i>	<i>996</i>
<i>Number of validations</i>	<i>10</i>
<i>Total number of datasets</i>	<i>3318</i>

This dataset classified into **9** medical specialities. Table 4.3 shows the Number of instances for each medical specialty.

**Table 4.3 Number of instances for each medical specialty**

<i>Cat.</i>	<i>Medical Specialty</i>	<i>Number of Reports</i>
<i>Cat:1</i>	<i>Cardiology</i>	<b>162</b>
<i>Cat:2</i>	<i>Emergency Medicine</i>	<b>407</b>
<i>Cat:3</i>	<i>Endocrinology</i>	<b>201</b>
<i>Cat:4</i>	<i>Family Practice/Primary Care</i>	<b>1446</b>
<i>Cat:5</i>	<i>Nephrology</i>	<b>64</b>
<i>Cat:6</i>	<i>Oncology</i>	<b>430</b>
<i>Cat:7</i>	<i>Otorhinolaryngology</i>	<b>330</b>
<i>Cat:8</i>	<i>Pulmonary Disease</i>	<b>171</b>
<i>Cat 9</i>	<i>Therapy, Physical</i>	<b>107</b>
<i>Total Number of Reports</i>		<b>3318</b>

#### **4.3.1 International Study in Asthma and Allergies in Childhood (ISAAC)at Grenada University**

Data for asthma patients collected by the University of Grenada from an asthma survey across Grenada in the West Indies. The results were provided by parents on behalf of Grenadian school children between the ages of 6 and 7 for the period October to December 2013. The survey was based on (ISAAC) study, which aids in cross-comparisons with other countries. Out of 2,362 surveys distributed, 1,374 were returned. However, only responses listing children’s ages between 6 and 8 years old were included in the analysis, resulting in 1,165 qualifying responses[98] .



The data collected included various aspects of information about asthmatic patients including age, symptoms, and other relatives with asthma, as well as the treatment suggested by the relevant doctors, as shown in Figure 4.4 A and B.

**Easy Breathing Survey**

Name: \_\_\_\_\_  
(Please Print Child's Name Clearly)

Today's Date:     /    /      
(Day/Month/Year)

1. What is your child's date of birth? \_\_\_\_\_  
(Day/Month/Year)
2. What is your child's gender? Male    Female
3. What parish do you live in? \_\_\_\_\_
4. Have your child experienced wheezing in the chest at any time in the last 12 months? Yes    No
5. Have your child awoken at night because of coughing in the last 12 months? Yes    No
6. Have your child experienced coughing, wheezing or shortness of breath with exercise or activity and had to stop because of these symptoms in the last 12 months? Yes    No
7. When your child has a cold, does the cough usually last for more than 10 days? Yes    No
8. Has a doctor ever diagnosed your child with asthma? Yes    No  
If yes, how old was your child when he/she was diagnosed with asthma? \_\_\_\_\_
9. If your child has been diagnosed with asthma, how many times in the past year has he/she:
  - a) Been hospitalized for asthma? \_\_\_\_\_
  - b) Had to visit a doctor because of asthma? \_\_\_\_\_
10. Is your child currently taking any medication for asthma? Yes    No
  - a) If yes, what medication(s)? \_\_\_\_\_
  - b) How often does your child use the medication? \_\_\_\_\_
11. How often does your child experience the following symptoms (please circle one of the answer choices):
 

a) Cough, wheeze or shortness of breath?	Less than 2x/week	More than 2x/week	Daily	Continuous
b) Difficulty Exercising	None	Rare	Occasional	Always
c) School absence due to symptoms?	None	Rare	Occasional	Frequent
12. If you think your child has asthma, how would you rate your severity of it? (Please circle 1 of the 5 choices)
  - I. Severe and extremely limits my activities
  - II. Moderate and limits some of my activities
  - III. Mild and rarely limits my activities
  - IV. Intermittent and does not bother me
  - V. I do not have asthma

*Figure 4.4 A, Patient survey for Grenada dataset.*

Breathe Grenada Survey # XXXX

12 Does anyone else in your family have asthma?	Yes	No
a) If yes, who? (In relation to your child).	<hr/>	
13. Is your child exposed to the any of the following:		
a) Cigarette smoke or any other forms of smoke?	Yes	No
b) Excessive Dust	Yes	No
c) Burning bush?	Yes	No
d) Pets at home?	Yes	No
e) Pollen	Yes	No
f) Landfills	Yes	No
14. Does your child experience difficulty breathing when exposed to the following:		
a) Cigarette smoke or any other forms of smoke?	Yes	No
b) Excessive Dust	Yes	No
c) Burning bush?	Yes	No
d) Pets at home?	Yes	No
e) Pollen	Yes	No
f) Landfills	Yes	No

*Figure 4.4 B, Patient survey for Grenada dataset.*

### **4.3.2 Asthma Dataset from Iraqi Hospitals**

A similar but local database was established by means of a survey conducted among a group of consultants specialising in chest and respiratory diseases at Imam Hussein medical city hospital in the Holy Karbala Governorate. The survey sheets were thus filled in by pulmonologist. Iraqi dataset consisted of 201 samples, 36 of which did not have the disease and 165 had the disease as shown in Figure 4.5 A and B.



**Easy Breathing Survey**

Name: \_\_\_\_\_  
(Please Print Child's Name Clearly)

Today's Date: \_\_\_\_/\_\_\_\_/\_\_\_\_  
(Day/Month/ Year)

1. What is your child's date of birth? 12/5/11  
(Day/ Month/ Year)
2. What is your child's gender? Male  Female
3. What parish do you live in? Karbala
4. Have your child experienced wheezing in the chest at any time in the last 12 months? Yes  No
5. Have your child awoken at night because of coughing in the last 12 months? Yes  No
6. Have your child experienced coughing, wheezing or shortness of breath with exercise or activity and had to stop because of these symptoms in the last 12 months? Yes  No
7. When your child has a cold, does the cough usually last for more than 10 days? Yes  No
8. Has a doctor ever diagnosed your child with asthma? Yes  No   
 If yes, how old was your child when he/she was diagnosed with asthma? 1 year
9. If your child has been diagnosed with asthma, how many times in the past year has he/she:
  - a) Been hospitalized for asthma? Never
  - b) Had to visit a doctor because of asthma? yes
10. Is your child currently taking any medication for asthma? Yes  No   
 a) If yes, what medication(s)? Butalin Symp.
- b) How often does your child use the medication? on need
11. How often does your child experience the following symptoms (please circle one of the answer choices):
  - a) Cough, wheeze or shortness of breath? Less than 2x/week  More than 2x/week  Daily  Continuous
  - b) Difficulty Exercising None  Rare  Occasional  Always
  - c) School absence due to symptoms? None  Rare  Occasional  Frequent
12. If you think your child has asthma, how would you rate your severity of it? (Please circle 1 of the 5 choices)
  - I. Severe and extremely limits my activities
  - II. Moderate and limits some of my activities
  - III. Mild and rarely limits my activities
  - IV. Intermittent and does not bother me
  - V. I do not have asthma

**Figure 4.5-A Iraqi Patient survey.**



12. Does anyone else in your family have asthma?	Yes	<input checked="" type="radio"/> No
a) If yes, who? (In relation to your child).	_____	
13. Is your child exposed to the any of the following:		
a) Cigarette smoke or any other forms of smoke?	<input checked="" type="radio"/> Yes	<input type="radio"/> No
b) Excessive Dust	Yes	<input checked="" type="radio"/> No
c) Burning bush?	Yes	<input checked="" type="radio"/> No
d) Pets at home?	Yes	<input checked="" type="radio"/> No
e) Pollen	Yes	<input checked="" type="radio"/> No
f) Landfills	Yes	<input checked="" type="radio"/> No
14. Does your child experience difficulty breathing when exposed to the following:		
a) Cigarette smoke or any other forms of smoke?	<input checked="" type="radio"/> Yes	<input type="radio"/> No
b) Excessive Dust	Yes	<input checked="" type="radio"/> No
c) Burning bush?	Yes	<input checked="" type="radio"/> No
d) Pets at home?	Yes	<input type="radio"/> No
e) Pollen	Yes	<input type="radio"/> No
f) Landfills	Yes	<input type="radio"/> No

**Figure 4.5 B Iraqi Patient survey.**

### 4.3.3 EMRs Patient Medical Reports Dataset

The Encounter dataset records interactions between a patient and healthcare providers for the purpose of providing healthcare services or assessing the health status of the patient. Data have been taken contained medical reports including symptoms and medication notes details of American patients in an emergency hospital collected during ten years. It is available online as a public, free-to-use dataset. It contains data about 3156 medical encounters, classified into 9 classes and 1176 medical fulfilments. Most of the data, however, is contained in the description column, free-text natural language description section. This column is thus used for transcription and the assignation of medical specialties. This column was thus assumed to contain the medical specialty for each case, which was then used as a label as shown in Figure 4.6.

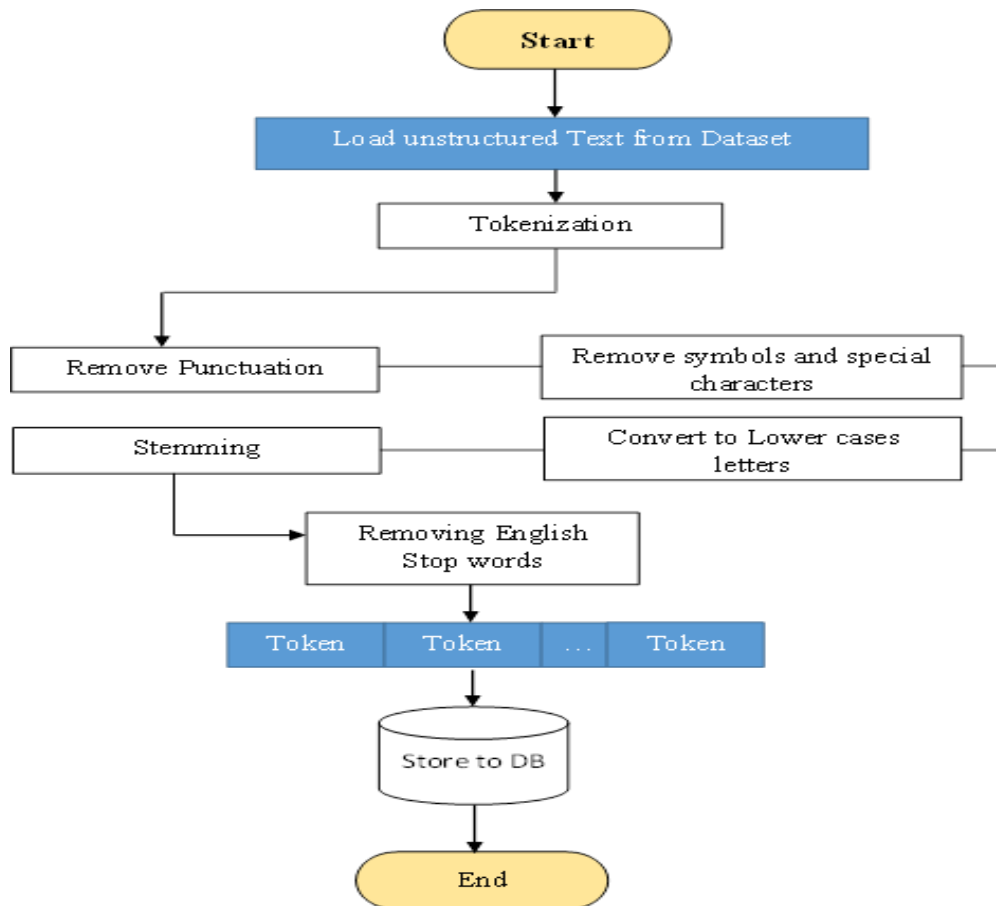
A1897		EMERGENCY MEDICINE
B	A	
pt reports severe weight loss_ moderate paresthesia in lower limbs. Also complains of severe increased thirst from time to time. he is a 60 YO male. o:Height 180 cm	EMERGENCY MEDICINE	1897
pt c/o 4 weeks h/o Hemorrhagic Stroke. NKDA. pt is a 48 YO F gastroenterologist. she denies smoking. patient has a history of excessive levels of alcohol consum	FAMILY PRACTICE/PRIMARY CARE	1898
48 yo female gastroenterologist presents with Hemorrhagic Stroke for 12 days. o:Height 60 in_ Weight 158 lbs_ Temperature 98.8 F_ Pulse 79_ SystolicBP 113_ D	CARDIOLOGY	1899
patient complains of severe weight loss_ severe b/l foot pain and moderate lethargy. NKDA. she is a white female aged 48 years. consumes more than 4 alcoholic	EMERGENCY MEDICINE	1900
female aged 62 years presents today for routine exam with history of severe b/l foot pain. patient complains of moderate lethargy_ moderate hyperesthesia in lower	FAMILY PRACTICE/PRIMARY CARE	1901
patient describes moderate hyperesthesia in lower limbs_ moderate lethargy and severe b/l foot pain. pt is a 62 YO female nanny. denies any alcohol use. Denies e	ENDOCRINOLOGY	1902
pt presents for exam. pt denies any specific issues. patient is a 62 yo white F. o:_ High-sensitivity fecal occult blood test Negative_ Bilateral Mammography no mass	ONCOLOGY	1903
pt presents without complaints. pt denies any issues. she is a 62 yo white f nanny. o:_ Intraocular Pressure eye pressure = 14 mmHg a:no complaints at this time	OTORHINOLARYNGOLOGY	1904
pt presents with progressive critical cough_ critical shortness of breath and critical dyspnea for past 3 weeks. she is a 62 yr old f. o:Height 166 cm_ Weight 43.6 kg	FAMILY PRACTICE/PRIMARY CARE	1905
pt presents with progressive critical cough_ critical shortness of breath and critical dyspnea for past 7 weeks. pt is a F nanny aged 62 years. o:Height 166 cm_ Wei	PULMONARY DISEASE	1906
patient presents with 3 years history of critical dyspnea. she is a white f aged 33 ys. o:Height 158 cm_ Weight 94.6 kg_ Temperature 36.9 C_ Pulse 85_ SystolicBP	EMERGENCY MEDICINE	1907
54 yo m locksmith c/o 7 weeks h/o Acute Renal Failure. NKDA. Patient reports that he never drinks alcohol. pt has a one pack per day habit. o:Height 71 in_ Weight	FAMILY PRACTICE/PRIMARY CARE	1908
male aged 54 Ys presents with mild difficulty breathing when laying down_ mild chest pain and palpitations and mild swollen ankles. he is having symptoms 3-5 x d	FAMILY PRACTICE/PRIMARY CARE	1909
female aged 28 yrs presents with Chronic Renal Failure for 6 weeks. denies any alcohol use. she smokes 1 pack a day. o:Height 156 cm_ Weight 62.4 kg_ Temper	FAMILY PRACTICE/PRIMARY CARE	1910
a 28 yo female c/o 10 days h/o Chronic Renal Failure. NKDA. denies any alcohol use. she smokes a pack/day for 2 years. o:Height 157 cm_ Weight 65 kg_ Tempe	NEPHROLOGY	1911
pt complains of severe b/l foot pain_ severe frequent urination and moderate hyperesthesia in lower limbs. NKDA. she is a 28 yo white F. denies any alcohol use. sh	EMERGENCY MEDICINE	1912
male nursemaid aged 40 ys presents with 13 months history of severe b/l foot pain_ moderate hyperesthesia in lower limbs. patient reports severe increased thirst.	EMERGENCY MEDICINE	1913
white male nursemaid aged 40 yrs presents for periodic physical. pt says he has no complaints today and no changes to PMH/PSH. Patient does not smoke. he re	FAMILY PRACTICE/PRIMARY CARE	1914
a 40 yo M nursemaid presents without complaints. patient denies any issues. o:_ Digital Rectal Exam no lump noted_ Skin cancer screening no abnormal skin or n	ONCOLOGY	1915
pt presents without specific complaints. patient denies any specific issues. he is a 40 yo white male. o:_ ECG normal rate and rhythm a:no current issues p:perform	CARDIOLOGY	1916
40 yr old white male presents for exam. pt denies any specific issues. o:_ Intraocular Pressure eye pressure = 14 mmHg a:no current issues p:performed Intraoc	OTORHINOLARYNGOLOGY	1917
a white male aged 40 yrs presents and denies any specific issues. o:_ Spirometry Fev1/FVC = 80% a:no complaints at this time p:performed Spirometry.	PULMONARY DISEASE	1918
patient C/O Chronic Renal Failure. NKDA. he is a male aged 41 ys. Patient does not smoke. occasional EtOH. o:Height 74 in_ Weight 135 lbs_ Temperature 99 F.	EMERGENCY MEDICINE	1919
41 yr old white M presents with Acute Renal Failure. he is having symptoms 2 x week_ in spite of treatment. he is a heavy drinker. he denies ever using cigarettes. c	FAMILY PRACTICE/PRIMARY CARE	1920

Figure 4.6 sample of EMR dataset.

## 4.4 Asthma Diagnosis using NLP & Data Mining

The methodology proposed in this model is depicted in Figure 4.1. Many steps were required to achieve the goals of this study:

- A pre-processing step, including tokenisation, stemming and normalisation, as shown in Figure 4.7.



*Figure 4.7 pre-processing steps*

- Feature extraction is used to extract features from text data include formats that are not accepted by machine learning techniques.
- Text classification for unstructured data to allow the database able to process text effectively.
- Applied different data mining algorithms to determine the best one based on comparing their results, as shown in Figure 4.1.

#### 4.4.1 Pre-processing stage

This stage was divided into two separate processes. The first one was the pre-processing of structured data, while the second was pre-processing of unstructured data. The latter consisted of multiple steps, as shown below.

##### 1) Pre-processing the structured data

###### *i. Nominal to Binary conversion*

In this step, the categorical attributes were transformed into a binary form

###### *ii. Multinomial to Numeric conversion*

This step was used to convert multinomial attributes such as “school absences due to symptoms” and “difficulty exercising”, measured by means of categorical values such as “Never, Rarely, Occasionally, always” into a numerical format (0, 1, 2, 3).

##### 2) Pre-processing the unstructured data

To achieve this effectively, multiple steps were required,

**1- Tokenisation.** In this study, tokenisation was performed by dividing up longer strings of text such as “Prescribed Medication”, “Relatives with asthma” into smaller pieces, or tokens (words), based on the spaces between them.

**2- Removal of punctuation, symbols, and special characters:** This step eliminated all unnecessary information carried by non-alphabetical or numeric characters such as '!', '"', '#', '\$', '%', '&', "'", '(, )', '\*', '+', ',', '-', '.', '/', ':', ';', '?', '[', '\\', ']', '\_', '^', '{', '|', '}', and '~', thus processing the text into an appropriate form.

###### **3- Removal of stop words**

This stage filtered out those words which contributed little or nothing to the overall meaning such as “the”, “to”, and “this”. To achieve this, an appropriate library was used to clean the data of all identified stop words.



**4- Stemming:** It is the process of removing unnecessary differential parts of words, such as suffixes and prefixes, that change the pronunciation and understanding of a word. Doing this offers in the origin of the word without any additions.

**5- Conversion to Lower Case**

This step converts all characters to lowercase, which simplifies the NLP task.

**All these tools in python are available in an open-source library called Natural Language Toolkit (NLTK).**

NLTK is a set of software modules, data sets, tutorials, and exercises that encompass the symbols and statistical processing of one or more natural languages. NLTK written in Python and distributed under an open license has become popular in recent years in teaching and research [99]. An NLTK provides a simple, uniform, and demonstrably extensible framework for assignments and projects. Those in Python are also thoroughly documented, easy to learn, and simple to use [100].

## **4.4.2 Feature Extraction and Weighting stage**

### **4.4.2.1 Feature Extraction**

The task of converting a particular text into a vector is among the most important steps in text processing, as this is required to extract the most important features from the text. Feature extraction methods are used to extract features from text data include formats that are not accepted by ML techniques; however, it is necessary for all features of the text to be used are extracted in a specific format accepted by these algorithms.

In this model, the TF-IDF feature extraction method was used, as this technique considers the number of times that a word appears across all documents in the document set. Mathematically, TF-IDF is the result of two scales: TF and IDF. Both TF and IDF values are necessary to calculate the TF-IDF, which arises from the idea that inverted document frequency is the statistic that most readily indicates the importance of a given word in a set of documents [51].

After the texts are converted using TF-IDF, and thus given appropriate weights, ML algorithms can be applied.

#### 4.4.2.2 Feature Weighting

##### *i. Creating a vector of features*

Unstructured text is not suitable for use in machine learning techniques directly. As part of the pre-processing stage, all punctuation, stop words and other irrelevant features are removed, and the FE technique selected then transforms the given text into a matrix appropriate for algorithmic use, in which rows represent the texts and columns represent the extracted features or words.

##### *ii. Calculating weights for features*

In this work, TF-IDF was employed to transform unstructured text into features, based on term frequency and occurrence. This technique considers the number of times that each term appears in all documents and across the document set. Once the TF and IDF values are obtained, using Eq. (3), it is possible to calculate the overall TF-IDF.

$$TF - IDF_i = t_{fi} \times \log \frac{N}{d_{fi}} \dots \dots \dots (1)$$

where  $t_{fi,j}$  is the number of times the term  $i$  appears in a document  $j$ ,  $N$  refers to the total number of documents, and  $d_{fi}$  is the number of documents that contain the term  $i$ .

##### *iii. Calculating weights for each instance*

While the previous step calculated weights for each extracted feature or word, in this step, the weight for each particular text is calculated using Eq. (4).

$$W_{i,j} = \sum_{j=0}^n TF - IDF_{i,j} \dots \dots \dots (2)$$

where  $w_{i,j}$  represents the weight for a particular text,  $n$  is the number of extracted features, and  $i,j$  = represent the rows and columns of the extracted matrix.

### **4.4.3 Text Mining Stage:**

Various AI techniques for text mining can be used to automatically process data and generate useful or valuable insights, enabling users to make decisions based on the information extracted from data. Text mining identifies facts, assertions, and relationships buried in a block of textual big data that without excavation may never have been discovered and which would thus have remained buried forever without anyone benefiting from them. Excavation allows extraction of this information, as well as its transformation into a structured model that can be either analysed further or presented directly[101].

In the field of data classification, several different techniques, such as RF and logistic regression methods, have been proposed. Although many other data mining techniques have been utilised in the previous literature, only the methods most relevant to this model are presented in this case, Support Vector Machine (SVM), Multiple Layer Perceptron (MLP) and DT.

## 4.5 Classification of Specialties in Textual Medical Reports Using Classical Algorithms

The methodology proposed in this model is depicted in Figure 4.1.

### 4.5.1 Data Pre-Processing

To achieve this effectively, multiple steps were required, as shown in Figure 4.9.

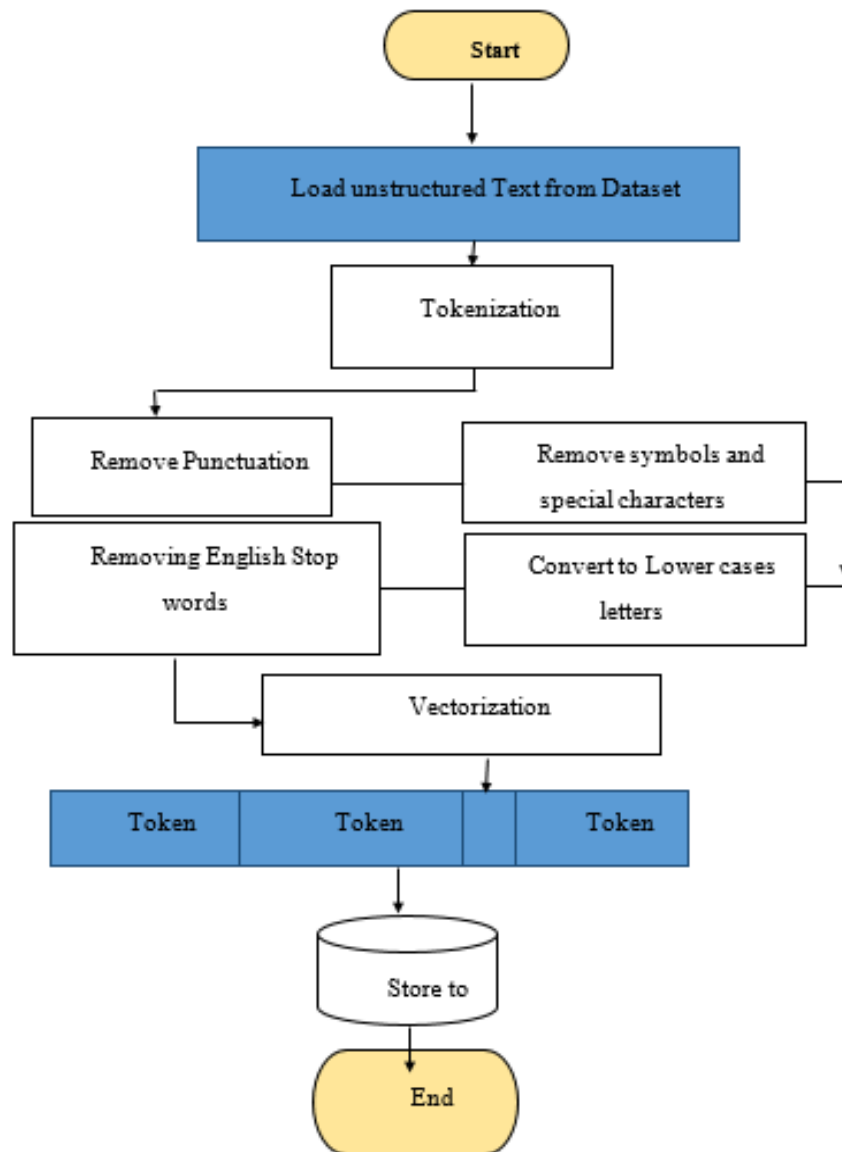


Figure 4.8 Block diagram of the pre-processing steps

The first stage of the proposed system is pre-processing. In this stage, data are preprocessed because the original dataset is not in the optimal format. Thus, they should be preprocessed before mining in order to clean the medical reports and the textual part in data contents from the noise and inappropriate information. Many different pre-processing methods were implemented in this thesis. The steps as shown below are used in the pre-processing of the medical texts.

#### **4.5.1.1 Word Tokenization**

Tokenisation was performed by dividing up longer strings of text such as “Prescribed Medication”, “Relatives with asthma” into smaller pieces, or tokens (words), based on the spaces between them. In this study, tokenization is performed by dividing the text document into words based on the spaces between them.

#### **4.5.1.2 Cleaning data**

In this step, the NLTK library in Python language was used to clean data from unnecessary information. This includes removing irrelevant information special characters, numbers, non-English letters, and elongated letters.

#### **4.5.1.3 Removing Punctuations**

The elimination of unnecessary information such as punctuations is a vital step in text pre-processing. Punctuations present how a sentence is created, how it should be read, and provide further explication for the meaning [102]. Examples of punctuations are: '!', '"', '#', '\$', '%', '&', "'", '(', ')', '\*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '\_', '~', '{', '|', '}', '~'.

#### 4.5.1.4 Removing Symbols and special characters.

Symbols in text t represent unnecessary information during the analysis stages.

Therefore, Examples of such symbols are ™ ́ ̀ ́ £ € ¥ ☺ ☻ ☼ ☽ ☾ ☿ ▲  
▶ ▼ ◀ ☺ ☻ ☼ ☽ ☾ ☿ ♀ ♂ ♠ ♣ ♥ ♪ # 🌿 ↗ ↘ ☪ ☸ ☹ ☺ ☻ ☼ ☽ ☾ ☿

#### 4.5.1.5 Removing Non-English Letters.

Step 5 shows the process of removing non-English letters from text. This includes all uppercase [A-Z] from A-to-Z letter and all lowercase [a-z] from a to z letter.

Removing Repeated and Elongated Letters.

#### 4.5.1.6 Removing Stop Words

The process of removing stop words. all stop words were removed from the medical text, this step filtered out those words which contributed little or nothing to the overall meaning such as “the”, “to”, and “this”. To achieve this, an appropriate library was used to clean the data of all identified stop words.

#### 4.5.1.7 Conversion to Lower Case

This step converts all characters to lowercase, which simplifies the NLP task. All these tools in python are available in an open-source library called (NLTK).

## **4.5.2 Feature Extraction (FE)**

After the pre-processing step, the task of converting a particular text into a vector space is an important part of text processing that can help extract the most important features from the text. FE methods are necessary to extract features as text data includes formats that are not accepted by ML techniques, and the features of such text must be extracted in a specific format that is accepted by such algorithms.

In this model, the TF-IDF feature extraction method was used.

### **4.5.2.1 Creating a Vector of Features**

Unstructured text not suitable for use directly by ML techniques. As the pre-processing stage removes all punctuation, stop words, and other irrelevant features, the role of FE techniques is to transform a particular text into a matrix appropriate for ML algorithms, with rows representing the texts and columns representing the extracted features or words.

### **4.5.2.2 Calculating Weights For Features.**

In this work, Term Frequency-TF-IDF was employed for transforming unstructured text into features based on term frequency and occurrence. This technique considers the number of times that a term appears in all documents and in the document set. Once the TF and IDF values are obtained according to Eq. (3), it is possible to calculate the TFIDF.



### **4.5.3 Feature Selection (FS) by using Chi-square Test**

Features selection based on important features, thousands of features can be found in text classification. Several types of feature selection techniques were tried and the best were chosen that led to the highest accuracy results. Accordingly, the choice of significant features was performed in this research based on the chi-square.

The chi-square test is a statistical test of independence to determine the dependency of two variables. It shares similarities with a coefficient of determination, the chi-square test is only applicable to categorical or nominal data. It is one of the simplest techniques for vocabulary reduction.

### **4.5.4 Data Classification**

In terms of such data classification, several different techniques, such as RF and LR methods, have been proposed. Although many other data mining techniques have been utilised in the literature, only the methods most relevant to this model were mentioned SVM, MLP and DT.

Text classification methods were used in order to classify 5447 medical text report into nine class. **Emergency Medicine, Otorhinolaryngology, Family Practice/Primary Care, Oncology, Physical Therapy, Nephrology, Pulmonary Disease, Cardiology, Endocrinology** These reports were fed into the pre-processing stage and selection stage to apply an appropriate classifier. The total number of datasets was 3318, contain number of sentences was 15153 divided into 2322 quantity as the training set and 996 quantities as the testing. with number of validation equal to 10 quantities.

## **4.6 Classification of specialties in textual medical reports by using deep learning**

This model has two main stages, preprocessing and CNN stage.

### **4.6.1 Data Pre-Processing**

In this stage, same as preprocessing in last model to clean data as shown below.

#### **4.6.1.1 Clean data**

To achieve this effectively, multiple steps were required to cleaning dataset before deep learning as shown below. output in this model was vector of words.

1. Stop words removal, stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc. The intuition behind using stop words is that, by removing low information words from text, we can focus on the important words instead.
2. Stemming is the process of reducing inflection in words to their root form. The “root” in this case may not be a real root word, but just a canonical form of the original word.  
Stemming uses a crude heuristic process that chops off the ends of words in the hope of correctly transforming words into their root form.
3. Lowercasing the text data is one of the simplest and most effective form of text preprocessing. It is applicable to most text mining and NLP problems and can help in cases when the dataset is not very large.

### **4.6.1.2 Encoding Techniques (Vector Representations of Text)**

In order to put the words into the ML algorithm, the text data should be converted into a vector representations.

In this step, by using a process called one-hot encoding, the categorical variables in text reports are converted into a numerical form.

The input to the most ML algorithms and all deep learning architectures algorithms must be binary values.

Every word which are part of the given text data are written in the form of vectors, constituting only of 1 and 0 .Each one hot vector being unique, therefore one hot vector is a vector whose elements are only 1 and 0. Each word is written or encoded as one hot vector, . This allows the word to be identified uniquely by it`s one hot vector so that no two words will have same hot vector representation.

### **4.6.2 Deep learning layers**

In this model five layers have been used as shown below:

#### **4.6.2.1 Word Embedded layer**

Embedding layer to transform the “one-hot vectors” into a sequence of dense vectors. Embedding Layer accepts a two-dimensional tensor of size  $S * L$  which is the encoded character sequence. Usually, the embedding layer is used for decreasing the dimension of the input tensor.

In this step zero-padding tool has been used to transform the input tensor into a fixed size.and solve the problem of one hot incoding in the previous stage, it can naturally treat the embedding layer as a look-up table.

After the preprocessing in embedding layer, the ConvNets can then extract the features by the convolutional kernel.

#### **4.6.2.2 Convolution neural network layer**

In convolutional layers, it can be applied up to three 1D-Convolution layers which have kernel size equal to five and feature map equal to 150 and used number of filters equal to 128 to extract 128 feature for each instance. The operation of convolution is widely used in text processing. For the 1D-Convolution that used, there are two signals which are text vector and kernel. After the process, the convolutional operation created a third signal which is the output. The text vector is the output from the embedding layer, vocabulary size equal to 15000 in our setting while kernel has length which is 5.

ReLU is a nonlinear activation function, which is widely employed in recent researches have been used which is unlike the 'sigmoid function, this activation function can handle the gradient vanishing problem better.

The threshold in ReLU can simulate the brain mechanism of human. L2 regulation is being used in all these layers because it is quite efficient for solving the overfitting problem.

#### **4.6.2.3 Pooling layer**

The max-pooling layer followed by the convolutional layer is necessary. The type of pooling layer used called max-pooling layer.

The pooling layer can select the most important features from the output of 1D-convolution. Also, it can diminish the parameters to accelerate the training speed. By choosing the temporal max-pooling with kernel size equal to the feature maps' number, only the most important feature remains in this stage.

#### **4.6.2.4 Flatten layer**

In this step data is converted into a 1-dimensional array for inputting it to the last dense layer. It has been flattened the output of the convolutional layers to create a single long feature vector. Then it is connected to the final classification model, which is called a fully-connected layer. The output of this stage was all the data in one line and make connections with the fully connected layer.

The features will be selected from pooling layer sent to the fully-connected layer with size  $128 * 1$  as a 1D tensor.

#### **4.6.2.5 Dense layer**

This layer is the most commonly used layer in artificial neural network. In any neural network, a dense layer is a layer that is deeply connected with its preceding layer which means the neurons of the layer are connected to every neuron of its preceding layer.

#### **4.6.2.6 The fully-connected layer**

Also known as the dense layer. At this stage, all the resulting features that selected from the max-pooling layer are combining.

The max-pooling layer selects feature from each convolutional kernel. The fully-connected layer can combine most of the useful assemble and then construct a hierarchical representation for the final stage, the output layer.

The output layer contain nine neurones because of the number of the target classes. Then apply the dropout layer and set the dropout rate to '0.1'.

Also the dense layer classifies the values emitted from the flatten layer. So it can experiment with different dimensions and input lengths in the embedding layer and different numbers of neurons in the dense layer to maximize accuracy.

## **4.6.3 Librares used in DL model**

### **4.6.3.1 TensorFlow**

TensorFlow is an open source ML framework for all developers. It is used for implementing ML and DL applications. In order to develop and research on fascinating ideas on AI, Google team created TensorFlow. TensorFlow is designed in Python programming language, hence it is considered an easy to understand framework.

### **4.6.3.2 Keras**

Keras is a DL application programming interface (API) written in Python, running on top of ML platform TensorFlow. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result as fast as possible is key for doing a fruitful research.

It is the high-level API of TensorFlow, an approachable, highly-productive interface for solving ML problems, with a focus on modern deep learning. It provides essential abstractions and building blocks for developing and shipping ML solutions with high iteration velocity.

Keras empowers engineers and researchers to take full advantage of the scalability and cross-platform capabilities of TensorFlow, being able to run Keras on Tensor Processing Unit (TPU) or on large clusters of Graphics Processing Units (GPUs), and you can export your Keras models to run in the browser or on a mobile device.

# CHAPTER 5

## RESULTS AND DISCUSSION

### 5.1 Introduction

This Chapter discusses the results of each stage for the proposed system as presented in Chapter four. The results of all stages are arranged based on their appearance in Chapter four. However, this Chapter begins with the hardware and software requirements in implementing the proposed system.

### 5.2 Software and Hardware

The proposed system was implemented using the following hardware and software requirements.

**Hardware:** Processor Intel i5, RAM 4GB, Storage 320 GB, Freq.1.7GHz,

**Memory:** 4096MB RAM.

**Software: Operating System:** Windows10 pro 64bit.

**Programming language:** Python language

**Libraries:** NLTK, TF-IDF, Tenser flow, kernel.

**IDLE:** The system was implemented by Python 3.7.3 shell, PyCharm.

Deep learning classifier implemented by using google Colab.





### 5.3.2 Feature Extraction Result

In this step, TF-IDF tool was applied on the three files of dataset to calculate the weight of each feature. The results are shown in figures 5.2, 5.3, 5.4 respectively.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	advair	aero	allegra	antibiotic	asthma	azithromy	becetide	becloasm: inhaler	becotech	becothind	birtal	bisclovan	brohec	bronor	
2		0	0	0	0	0	0	0	0.547319	0	0	0	0	0	0	0
3		1	0	0	0	0	0	0	0.589383	0	0	0	0	0	0	0
4		2	0	0	0.798812	0	0	0	0	0	0	0	0	0	0	0
5		3	0	0	0.514626	0	0	0	0	0	0	0	0	0	0	0
6		4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7		5	0	0	0	0	0	0	0	0	0	0.881722	0	0	0	0
8		6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9		7	0	0	0	0	0	0	0	0.40322	0	0	0	0	0	0
10		8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11		9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12		10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13		11	0	0	0	0	0	0	0	0	0.914338	0	0	0	0	0
14		12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15		13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16		14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17		15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18		16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19		17	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5.2 file of medication column after feature extraction

	Q	P	O	N	M	L	K	J	I	H	G	F	E
		nephew	nieces	neice	mother	randfathe	father	daughter	cousin	brothers	brother	sister	
		0	0	0	0	0	0.7608	0	0	0	0	0	
		0	0	0	0	0	0	0	0.60088	0	0	0	
		0	0	0	0	0.50018	0	0	0	0	0	0	
		0	0	0	0	0	0	0	0	0	0	0	0
		0	0	0	0	0	0	0	0	0	0	0.486	
		0.5918	0	0	0	0	0.54094	0	0	0	0	0	
		0	0	0.50018	0	0	0	0	0.64243	0	0	0	
		0	0	0	0	0	0	0	0.65687	0	0	0	
		0	0	0	0	0	0.60088	0	0	0	0	0	
		0	0	0	0.77491	0	0	0.63207	0	0	0	0	
		0	0	0	0	0	0	0	0	0	0	0	
		0	0	0	0	0	0	0	0	0	0	0	
		0	0	0	0.53204	0	0.65441	0	0	0	0	0	
		0	0	0	0	0	0	0	0	0	0	0	
		0	0	0	0	0	0	0	0	0	0	0	

Figure 5.3 file of relative column after feature extraction

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	age	gender	4.wheezin	if_1.how_of	Been_hosj	anyone_ah	a.child-exj	b.child-exj	c.child-exj	d.child-exj	e.child-exj	f.child-exj	a	b.difficulty	c.difficulty	d.difficulty	e.difficulty	f.difficulty	Difficulty	School_ab	SEVERITY	target
2	16	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	16	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	4
4	13	2	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
5	11	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	11	1	1	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
7	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	10	2	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1
9	8	1	1	0.66	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	2	2
10	8	2	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	8	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
12	8	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	8	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
14	8	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
15	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	8	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
18	8	2	1	5	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
19	8	2	1	5	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2
20	8	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
21	8	2	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	8	1	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
26	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	8	1	1	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	1

Figure 5.4 file of structure columns after feature extraction

### 5.3.3 Classifier Result

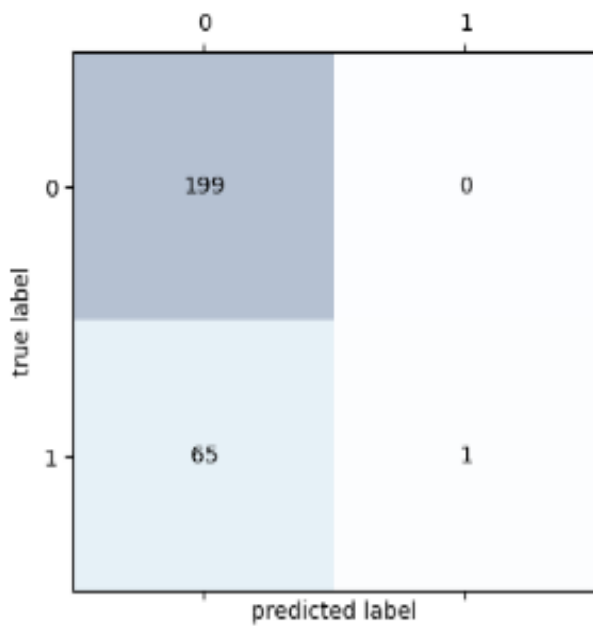
ML algorithms were applied on the structured part of dataset without NLP step. The results of the three algorithms are shown in table 5.1

*Table 5.1 The Performance Metric of the Three Algorithms without NLP technique (10 cross validation)*

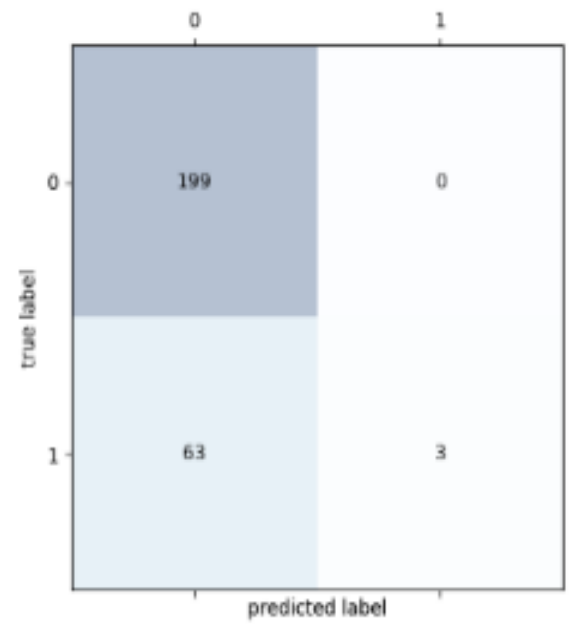
<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<i>75.07</i>	<i>42.88</i>	<i>37.53</i>	<i>50.01</i>
<i>ML-Perceptron</i>	<u><i>76.58</i></u>	<u><i>49.27</i></u>	<u><i>85.35</i></u>	<u><i>53.15</i></u>
<i>DT</i>	<i>75.44</i>	<i>44.46</i>	<i>87.67</i>	<i>50.75</i>

From the above results it can be seen that the accuracy was low for the structured data part only because the information of the unstructured part was neglected as the algorithms cannot deal with unstructured data. That leads to the inability to classify the diseased cases properly.

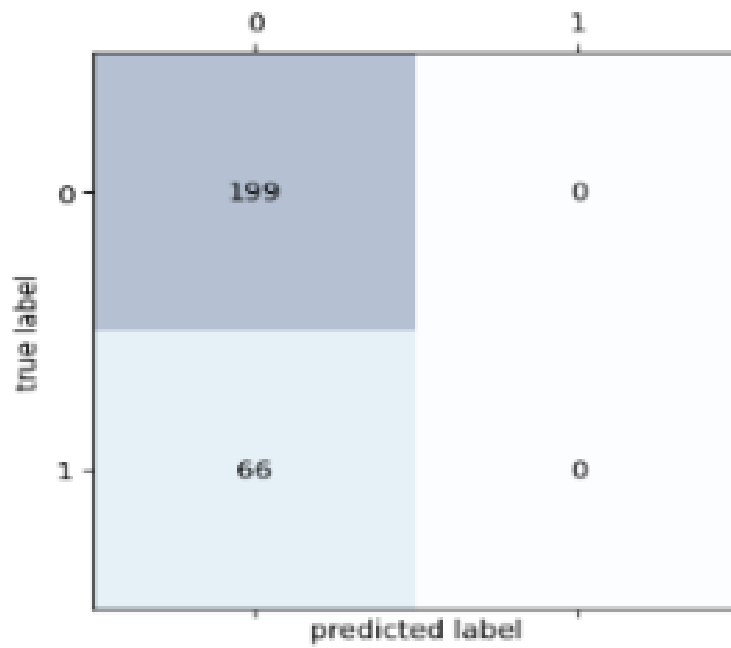
The confusion matrix for each algorithm is shown in Figure 5.5



*DT*



*MLP*



*SVC*

*Figure 5.5 The Confusion Matrix of the Three Algorithms for only structured Grenada dataset*

Then the model is applied on semi-structured dataset with NLP steps. The improvement in accuracy of three algorithms shown in table 5.2.

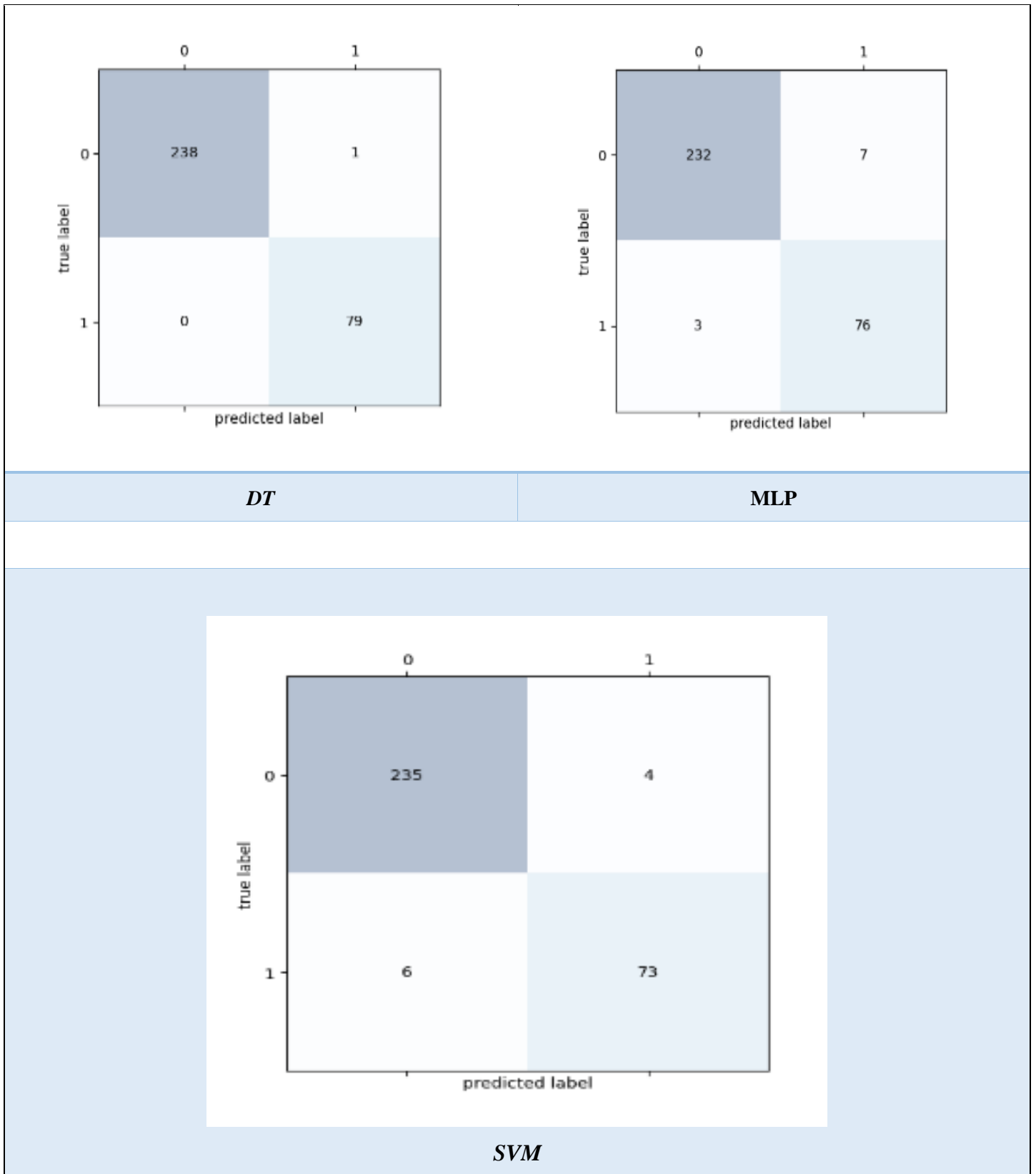
**Table 5.2 The Performance Metric of the Three Algorithms without Chi- FS technique (10 cross validation)**

<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>				
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>	<i>time</i>
<i>SVC</i>	<i>96.69</i>	<i>95.55</i>	<i>95.85</i>	<i>95.26</i>	<i>0.02 seconds</i>
<i>ML-Perceptron</i>	<u><i>97.54</i></u>	<u><i>96.72</i></u>	<u><i>96.72</i></u>	<u><i>96.72</i></u>	<u><i>0.57 seconds</i></u>
<i>DT</i>	<i>85.26</i>	<i>79.04</i>	<i>81.38</i>	<i>77.41</i>	<i>0.001 seconds</i>

In the table above, it can be seen that there is a significant change in the results after adding the data extracted from the unstructured columns, as it had a significant impact on improving the results since the algorithms extract more information from the added columns.

The highest accuracy obtained when applying MLP algorithm.

Figure 5.6 shows the confusion matrix for each algorithm when applied on Semi-structured dataset with NLP steps.



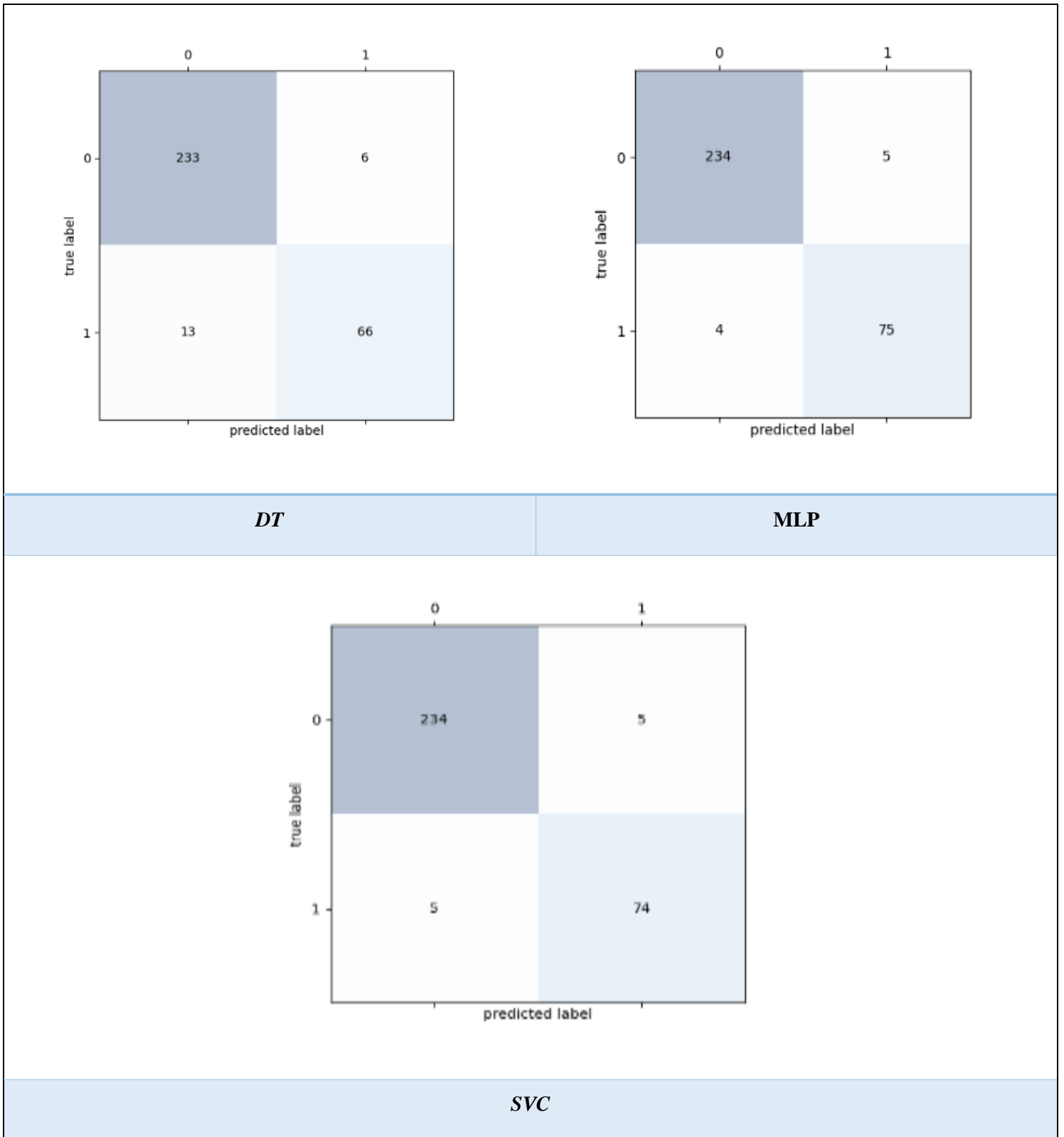
**Figure 5.6** The Confusion Matrix of the Three Algorithms for Semi Structured Grenada dataset without FS technique

Then Feature selection technique (Chi-square) was employed to enhance the performance of the proposed model and applied on semi-structured (Grenada-Iraqi) datasets with NLP steps. The improvement in accuracy of three algorithms shown in table 5.3.

**Table 5.3 The Performance Metric of the Three Algorithms with FS technique (10 cross validation)**

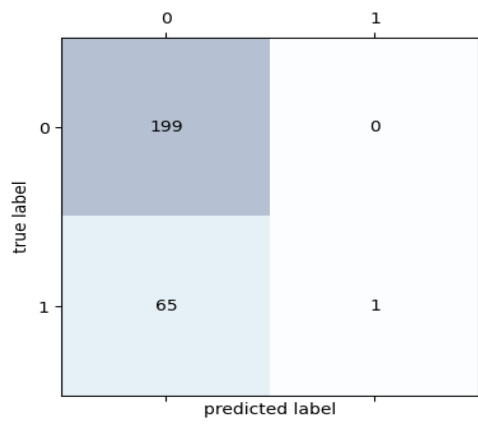
<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>				
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>	<i>time</i>
<i>SVC</i>	<i>96.88</i>	<i>95.82</i>	<i>95.99</i>	<i>95.64</i>	<i>0.02 seconds</i>
<i>ML-Perceptron</i>	<u><i>97.73</i></u>	<u><i>96.94</i></u>	<u><i>97.44</i></u>	<u><i>96.46</i></u>	<u><i>0.48 seconds</i></u>
<i>DT</i>	<i>89.42</i>	<i>85.49</i>	<i>86.52</i>	<i>84.60</i>	<i>0.001 seconds</i>

According to the above table, the positive impact on the accuracy of the results can be observed by adding the feature selection to the achieved results. Several methods were tried to choose the feature, and the chi-square method was the best and fastest method and the highest accuracy obtained when MLP algorithm applied.

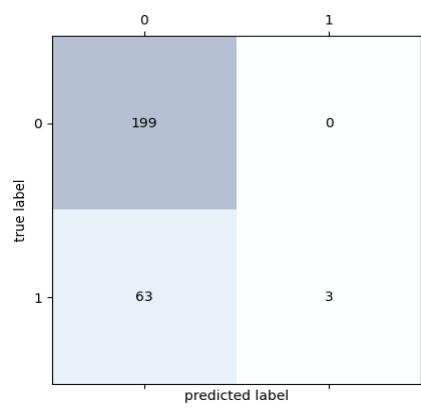


**Figure 5.7 The Confusion Matrix of the Three Algorithms for Semi Structured Grenada dataset with FS technique**

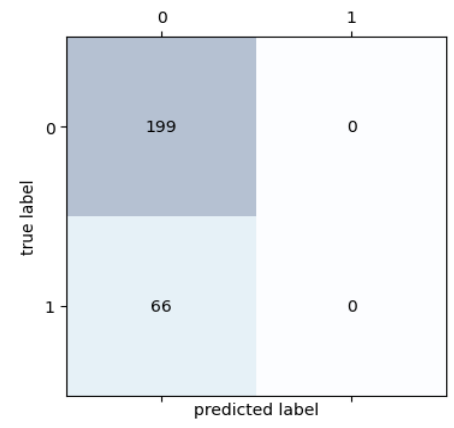
Figure 5.8 shows the improvement in results when NLP and FS techniques are employed with the proposed model.



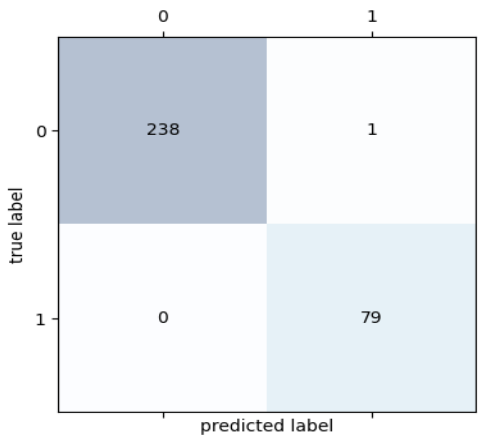
***DT without NLP***



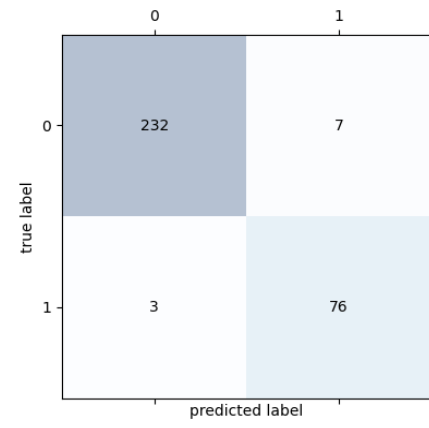
***MLP without NLP***



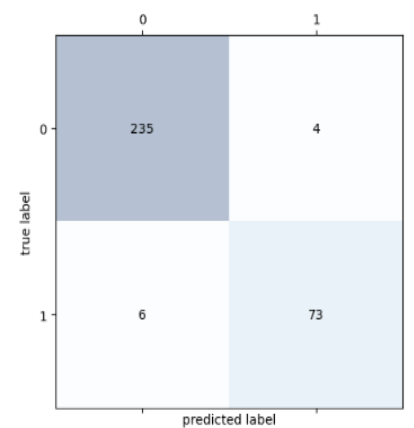
***SVC without NLP***



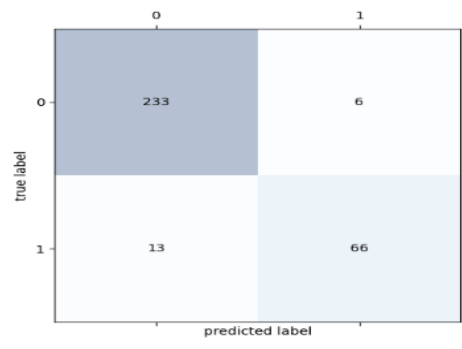
***DT with (NLP)***



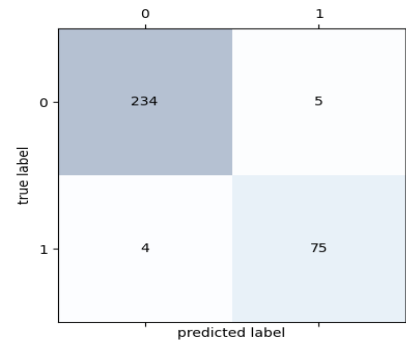
***MLP with (NLP)***



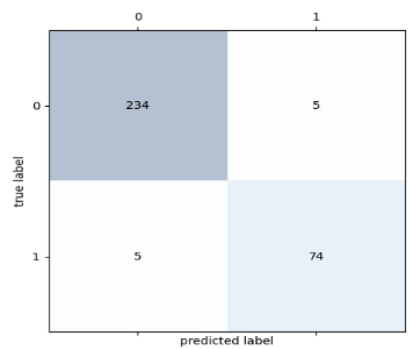
***SVC with (NLP)***



***DT with (NLP and FS)***



***MLP with (NLP and FS)***



***SVC with (NLP and FS)***

***Figure 5.8 The Confusion Matrix of the Three Algorithms for Semi Structured Grenada dataset.***



The previous three models were applied to the Iraqi local dataset, and the results for each model are shown in the tables 5.4,5.5 and 5.6, respectively.

**Table 5.4 The Performance Metric of the Three Algorithms without NLP technique for Iraqi dataset (10 cross validation)**

<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<b>81.59</b>	<b>44.93</b>	<b>40.79</b>	<b>50.00</b>
<i>ML-Perceptron</i>	<b><u>81.09</u></b>	<b><u>44.78</u></b>	<b><u>40.75</u></b>	<b><u>49.69</u></b>
<i>DT</i>	<b>79.10</b>	<b>44.16</b>	<b>40.56</b>	<b>48.47</b>

As shown in the table above the results without using NLP were low in accuracy and all the other metrics because of the little information available in addition to the little number of the samples used.

**Table 5.5 The Performance Metric of the Three Algorithms with NLP technique for Iraqi dataset and without FS (10 cross validation)**

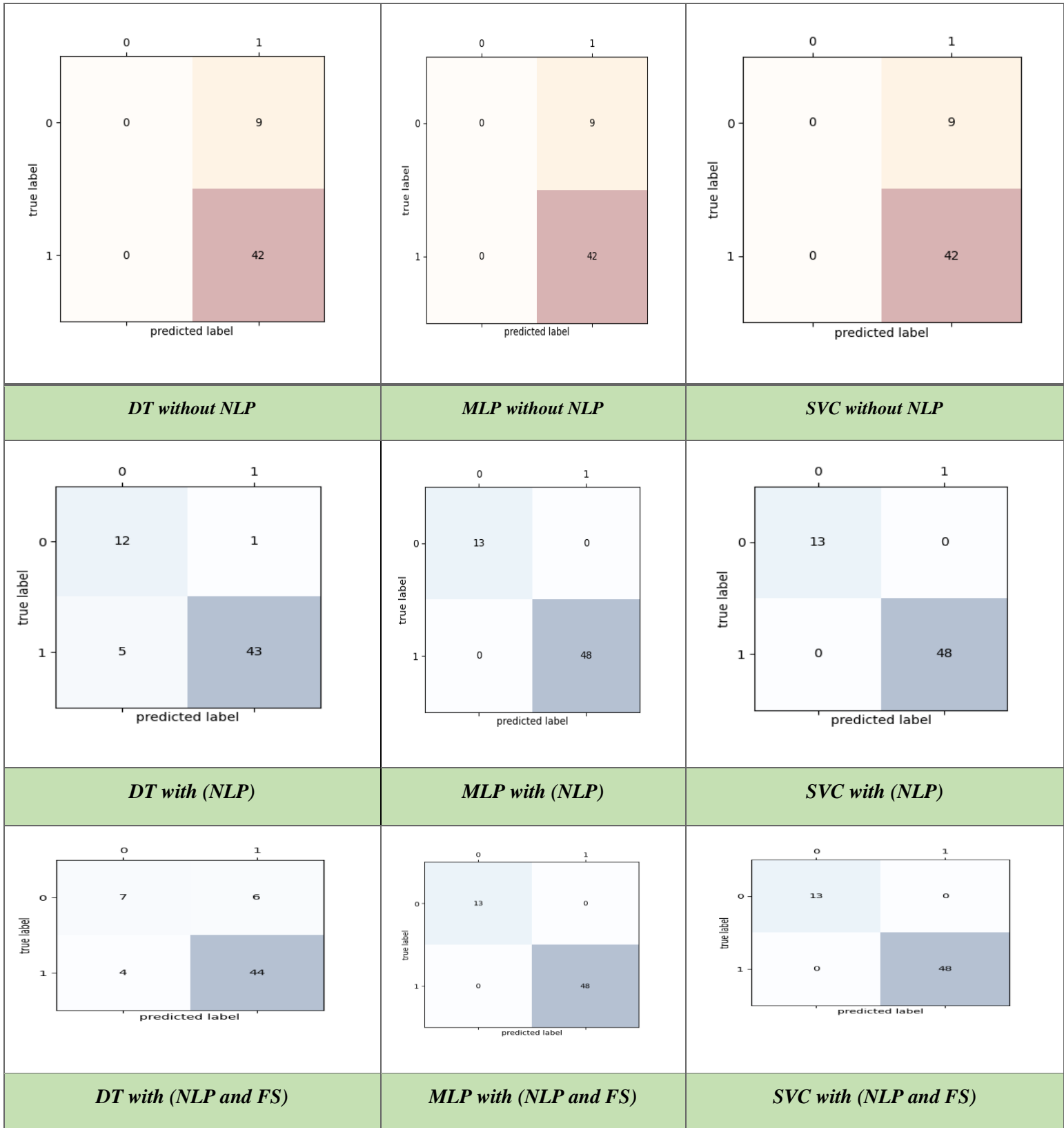
<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<b>98.00</b>	<b>96.68</b>	<b>96.68</b>	<b>96.68</b>
<i>ML-Perceptron</i>	<b><u>98.00</u></b>	<b><u>96.61</u></b>	<b><u>97.66</u></b>	<b><u>95.64</u></b>
<i>DT</i>	<b>85.57</b>	<b>78.73</b>	<b>76.35</b>	<b>82.78</b>

As shown in the above table applying NLP on the same dataset used previously leads to improving of all the metrics and the higher accuracy obtained when MLP applied while the higher recall was with SVC algorithm.

**Table 5.6 The Performance Metric of the Three Algorithms with NLP technique for Iraqi dataset and with FS (10 cross validation)**

<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<b>98.50</b>	<b>97.48</b>	<b>98.00</b>	<b>96.99</b>
<i>ML-Perceptron</i>	<b><u>98.81</u></b>	<b><u>98.30</u></b>	<b><u>99.39</u></b>	<b><u>97.29</u></b>
<i>DT</i>	<b>90.54</b>	<b>84.42</b>	<b>84.08</b>	<b>84.78</b>

In the table above the addition of FS technique on the same data leads to improving the accuracy because that FS chose only the most important features.



***Figure 5.9 The Confusion Matrix of the Three Algorithms for Semi Structured Iraqi dataset***

Figure 5.9 shows the improvement in results when NLP and FS techniques are employed with the proposed model.

## 5.4 Results of Classical classification model

### 5.4.1 Results of Data Pre-processing

The results of the pre-processing stage are shown after performing word tokenization to split words, cleaning medical text report by removing irrelevant information. The outcomes of the pre-processing stage were a collection of tokens for each text.

#### 1. Word Tokenization

Table 5.7 shows the results of splitting texts of EMR into its constituent words.

*Table 5.7 Word Tokenization*

<b>Before Tokenization</b>	:52-year-old white M lawyer presents without specific complaints. pt. denies any specific issues. o:, Intraocular Pressure eye pressure = 14 mmHg, Oral exam no sores or red and white patches a:no complaints at this time p:performed Intraocular Pressure, Oral exam.
<b>After Tokenization</b>	, : , 52 , - , year , - , old , white , M, lawyer , presents , without , specific , con specific , issues. O, : , ,Intraocular , Pressure , eye , pressure , = , 14 , mmHg, Oral , exam , no , sores , or , red , and , white , patches, a , : , no , complaints , at , this , time , p , : , performed , Intraocular , Pressure, Oral , exam. ,

In this step the texts are tokened into constituent words separated by commas.

#### 2. Removing Punctuations

Punctuations were removed from all text reports. Table 5.8 demonstrates a sample of the report before and after removing punctuations.

**Table 5.8 Removing Punctuation**

<b>Before Removing Punctuation</b>	, : , 52 , - , year , - , old , white , M, lawyer, presents , without , specific , complaints , , pt. denies any specific issues. O, : , ,Intraocular , Pressure , eye , pressure , = ,14, mmHg, Oral , exam , no , sores , or , red , and , white , patches, a:no , , complaints , at , this , time , p ,: , performed , Intraocular , Pressure, Oral , exam.
<b>After Removing Punctuation</b>	, 52 , year , old , white , M, lawyer, presents , without , specific , complaints , , denies any specific issues. ,Intraocular , Pressure , eye , pressure, mmHg, Oral , exam , no , sores , or , red , and , white , patches, a:no , complaints , at , this , time, performed , Intraocular , Pressure, Oral , exam. ,

Result of preprocessing step after cleaning data was as shown in table 5.9.

**Table 5.9 Result of Preprocess step on the text of the EMR**

NO.	EMR Sample
1	['yr', 'old', 'female', 'crystalographer', 'presents', 'today', 'routine', 'exam', 'patient', 'reports', 'acute', 'problems', 'patient', 'reports', 'never', 'drinks', 'alcohol', 'denies', 'smoking', 'oheight', 'cm', 'weight', 'kg', 'temperature', 'pulse', 'systolicbp', 'diastolicbp', 'respiration', 'hpv', 'ih', 'risk', 'dna', 'probe', 'negative', 'hpv', 'visual', 'acuity', 'study', 'right', 'eye', 'left', 'eye', 'anormal', 'exam', 'curent', 'issues', 'problem', 'status', 'hypertension', 'managed', 'administered', 'imunization', 'fluarix', 'pcal', 'ofice', 'reaction', 'imunization', 'fup', 'one', 'year', 'anual', 'checkup', 'soner', 'new', 'symptomsproblems', 'arise']
2	['sa', 'year', 'old', 'presents', 'critical', 'dyspnea', 'critical', 'shortnes', 'breath', 'critical', 'cough', 'patient', 'reports', 'never', 'drinks', 'alcohol', 'patient', 'one', 'pack', 'per', 'day', 'habit', 'oheight', 'cm', 'weight', 'kg', 'temperature', 'pulse', 'systolicbp', 'diastolicbp', 'respiration', 'fev', 'fev', 'fevfvc', 'fevfvc', 'arterial', 'blod', 'gas', 'paco', 'mhgpao', 'mhg', 'plum', 'acesory', 'muscle', 'use', 'heart', 'normal', 'murmurs', 'hent', 'wnl', 'rales', 'bil', 'achronic', 'obstructive', 'pulmonary', 'disease', 'padministered', 'oms', 'via', 'nasal', 'canula', 'contin', 'performed', 'fev', 'fevfvc', 'arterial', 'blod', 'gas', 'performed', 'emergency', 'services', 'level', 'completed']
3	['sa', 'white', 'female', 'aged', 'ys', 'presents', 'months', 'history', 'mild', 'spels', 'vertigo', 'pt', 'also', 'reports', 'increased', 'frequency', 'mild', 'ringing', 'ears', 'mild', 'headaches', 'particularly', 'back', 'head', 'morning', 'oheight', 'cm', 'weight', 'kg', 'temperature', 'pulse', 'systolicbp', 'diastolicbp', 'respiration', 'heart', 'systolic', 'murmur', 'base', 'heart', 'chest', 'clear', 'auscultation', 'bl', 'rales', 'whezing', 'extremities', 'edema', 'clubing', 'heart', 'normal', 'ahypertension', 'performed', 'level', 'established', 'patient', 'completed', 'prescribed', 'hydrochlorothiazide', 'mg', 'po', 'qd', 'ordered', 'basic', 'metabolic', 'panel', 'lipid', 'panel']

In the above table the resulted data from the preprocessed step was a set of words separated by commas can be dealt with as a vector.

### 5.4.2 Results of Features Extraction

Features are obtained based on the notion of the (TF-IDF). Each text was converted into a vector of features term frequency-inverse document frequency (TF-IDF). The outcome of this step was a vector of features.

### 5.4.3 Result of classification algorithms

After applying the model to classify the dataset into (9) multiclass without flowing through the FS pipeline to explain the effectiveness of FS on the classification results. Table 5.10 shows these results.

*Table 5.10 The Performance Metric of the Three Algorithms without FS technique (10 cross validation)*

<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<i>98.04</i>	<i>96.22</i>	<i>98.62</i>	<i>94.35</i>
<i>ML-Perceptron</i>	<u><i>98.82</i></u>	<u><i>98.18</i></u>	<u><i>98.52</i></u>	<u><i>97.88</i></u>
<i>DT</i>	<i>97.89</i>	<i>96.18</i>	<i>95.91</i>	<i>96.49</i>

In the above table the highest accuracy was obtained when MLP applied before FS step.

Then repeating the test on the same data with the addition of FS the result shown below in table (5.11).

**Table 5.11 The Performance Metric of the Three Algorithms with Chi- FS technique (10 cross validation)**

<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<i>98.22</i>	<i>96.86</i>	<i>98.68</i>	<i>95.39</i>
<i>ML-Perceptron</i>	<u><i>98.81</i></u>	<u><i>98.63</i></u>	<u><i>98.81</i></u>	<u><i>98.46</i></u>
<i>DT</i>	<i>98.07</i>	<i>96.66</i>	<i>96.51</i>	<i>96.83</i>

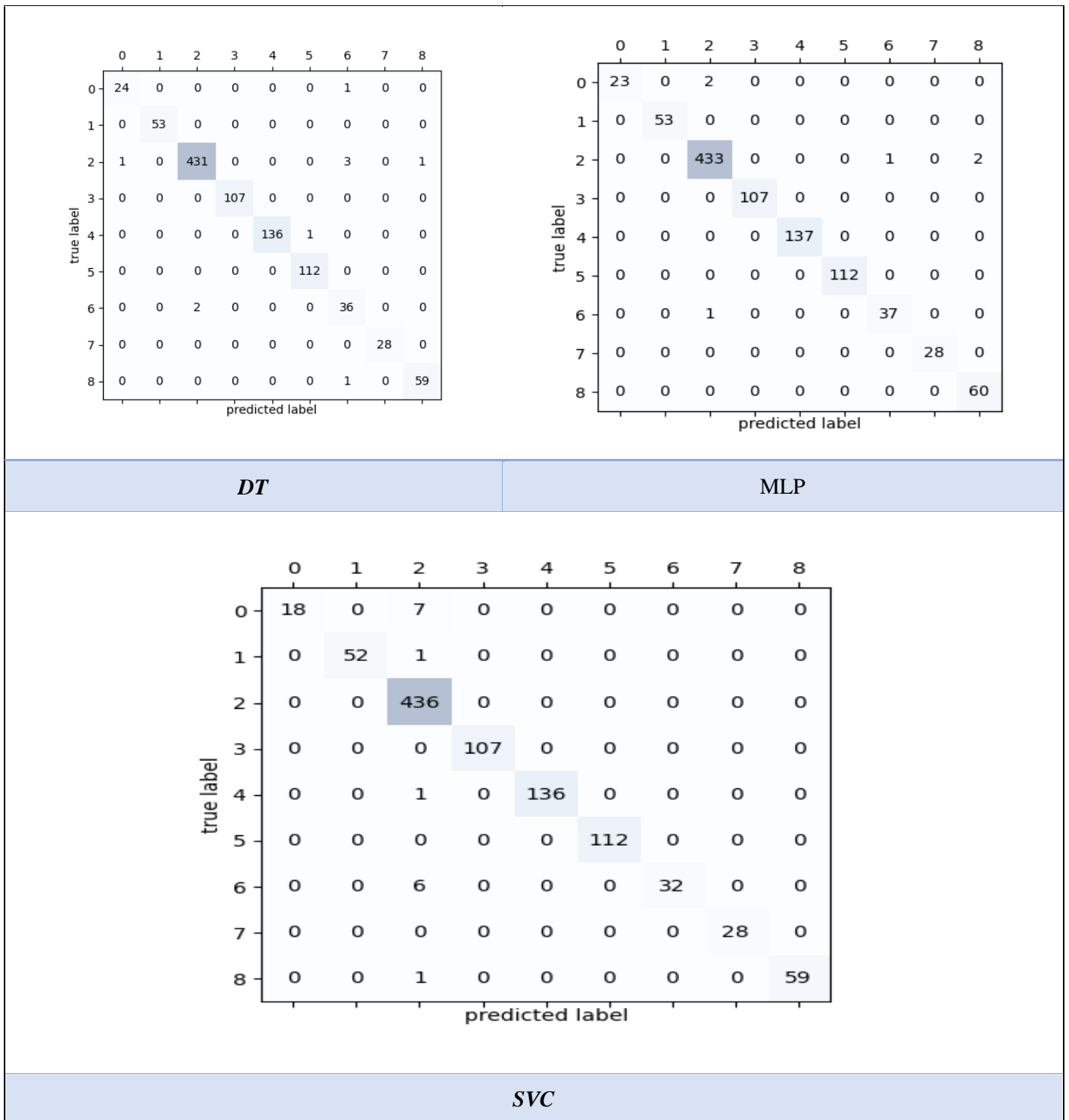
In this step after applying several techniques of feature selection technique (FS), Chi- FS technique gives the best result in this case. The effect of FS is improving the accuracy, and the best algorithm was MLP algorithm.

Table 5.12 showed the execution time of the three algorithms. MLP have the greatest impact on time after the addition FS. The added of FS was reducing execution time of the three algorithms and this effect is very important in classification. DT have the least time of execution.

**Table 5.12 The Execution Time of the Three Algorithms**

<i>Text Mining Algorithm</i>	<i>Execution Time (Second)</i>	
	<i>Without FS</i>	<i>With FS</i>
<i>SVC</i>	<i>3.5 seconds</i>	<i>3.19 seconds</i>
<i>ML-Perceptron</i>	<i>18.22 seconds</i>	<i>15.86 seconds</i>
<i>DT</i>	<i>0.57 seconds</i>	<i>0.53 seconds</i>

The Confusion Matrix shows the results of the three algorithms with Chi\_FS technique as illustrated in Figure 5.10.



**Figure 5.10 The Confusion Matrix of the Three Algorithms with Chi & FS Technique**



## 5.5 Result of deep learning CNN classifier

### 5.5.1 Preprocessing results

In this step data collected from EMRs and starting the steps of clean data to remove unwanted feature in medical report to clean report then vectorize the features to be as a vector able to encoding, then one hot encoding decodes all features in this vector, the result of this step was vector of encoding features.

### 5.5.2 Embedding results

In this step. The received vector of feature from the previous step have several lengths for each report. Padding tool applied to solve this problem by adding zeroes to make reports have the same length of features. Then word embedded work to aggregate the convergences feature in same place in matrix. The result of this step was two-dimension matrix.

### 5.5.3 CNN results

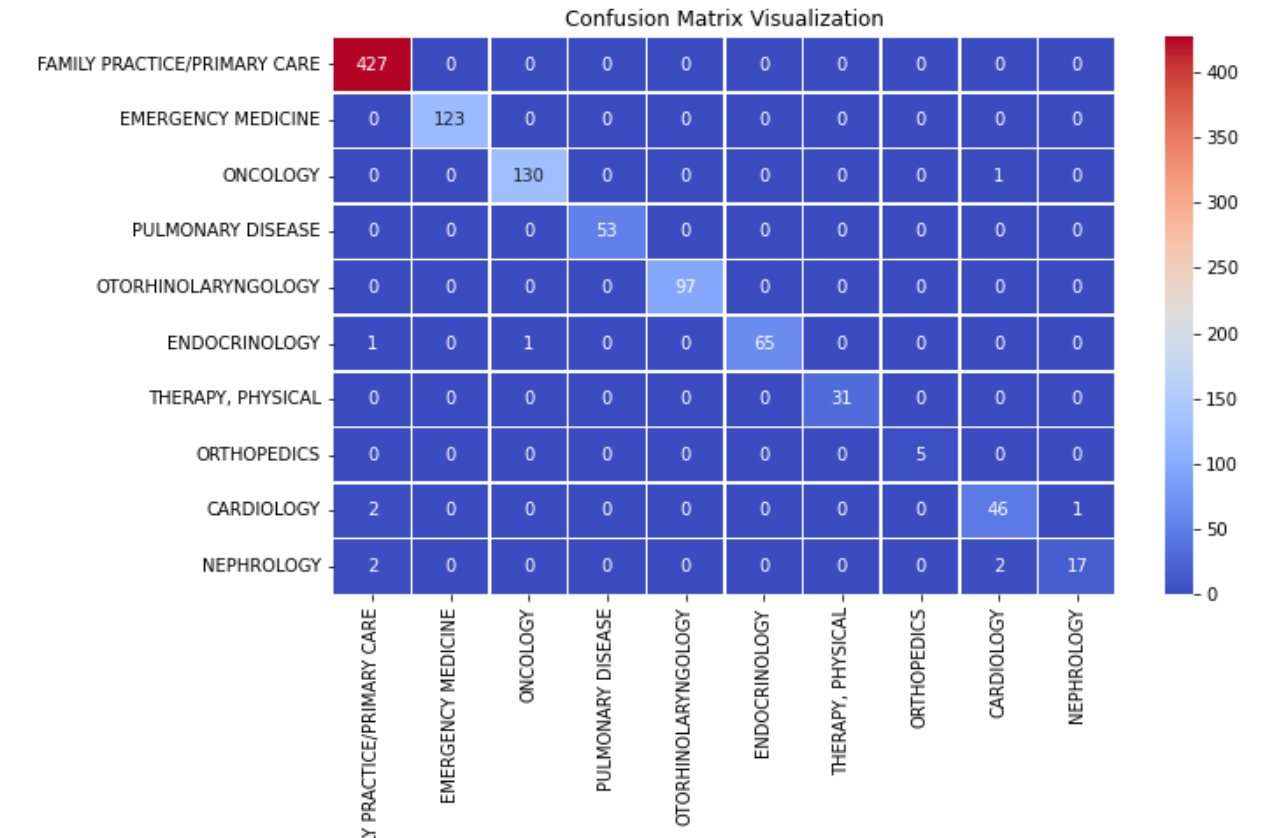
The purpose of this step was classifying the output of flatten layer to ten classes as an output of this model, the result was shows in table 5.13.

This model was implemented using the (**Colab**) website of Google to take advantage of the libraries stored automatically on this site without the need to download them to the user's computer.

*Table 5.13 deep learning results.*

Accuracy	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>	<i>time</i>
<b>99.00</b>	<b>97.82</b>	<b>98.64</b>	<b>97.11</b>	<b>2m&amp;16seconds</b>

Figure 5.11, below shows the confusion matrix for the model explain the result of classifier on the reports into ten classes.



**FIGURE 5.11** deep learning CNN results

Table 5.14 compares between the result of classical classification in the model in Figure 4.2 and the result of deep learning CNN results.

*Table 5.14 Comparison of classical classification  $r$  and deep learning CNN results*

<i>Text Mining Algorithm</i>	<i>Accuracy Metrics</i>			
	<i>Accuracy</i>	<i>F1-Measure</i>	<i>precision</i>	<i>Recall</i>
<i>SVC</i>	<b>98.22</b>	<b>96.86</b>	<b>98.68</b>	<b>95.39</b>
<i>ML-Perceptron</i>	<b><u>98.81</u></b>	<b><u>98.63</u></b>	<b><u>98.81</u></b>	<b><u>98.46</u></b>
<i>DT</i>	<b>98.07</b>	<b>96.66</b>	<b>96.51</b>	<b>96.83</b>
<b><i>CNN DEEP LEARNING</i></b>	<b>99.00</b>	<b>97.82</b>	<b>98.64</b>	<b>97.11</b>

## 5.6 Discussion

In the first model for asthma diagnosis, the proposed model with the addition of TF-IDF has proven effective in identifying the best features of data and removing inappropriate or negative influences on the part of unstructured data from semi structured dataset based on the use of NLP techniques. Overall, this offered promising results for all the algorithms examined. By comparing the results of the data generated by NLP tools with the results of the original data, it can be noted that an increase in the accuracy index and other indicators, which have a significant role in increasing the accuracy of diagnosis. The greatest increase in F1-measure, precision, and F1-measure. Precision was seen for ML-Perceptron, at 98.86%. 98.86%, and 98.87%, respectively, for the Grenada University data. For the Iraqi data, ML-Perceptron was the best algorithm in terms of accuracy. precision metrics. Here, F1-measure and precision and F1-measure. precision was equal to 97.51%, 95.72%, and 97.331%, respectively. These are very high rates that have not been reached in the researches similar to this work.

In the second model, the results obtained after applying AI techniques (NLP with TF-IDF), without feature selection, accuracy was reached (98%) according to table 5.3 due to the use of pre-processing techniques, represented by NLP steps. A FS step was added to reduce the number of features, retaining only the most effective features. The addition of this step significantly contributed to increase the accuracy of the implementation of all the algorithms used, as shown in table (5.3). Another important factor in reducing the training and testing time is shown in table (5.4), the training time of the algorithms with (Chi-Square) FS was reduced significantly. The highest accuracy was achieved by applying MLP classification techniques (98.81%).

**Table 5.15** shows the results for some of related work in this field where researchers have tried to extract medical information from a database by using several forms of pre-processing and mining.

**Table 5.15 Methods and result for classification related work**

<i>No.</i>	<i>Research paper</i>	<i>Type of data</i>	<i>Mining method</i>	<i>Result</i>	<i>Country and year</i>
1	Li et al. [22]	EHR. MIMIC-III dataset	LSTM and CNN-based models	0.3464 on AUCROC	USA 2021
				Improves 0.4521 on F1.	
2	Thomas et al. [35].	18,453 pathology reports on a prostate cancer	Latent Dirichlet Allocation (LDA) +Red-LDA	99.1 % sensitivity	Dublin, Ireland 2014
				99.9 % specificity	
				97.6 %. overall ability	
3	Weng et al. [33]	(iDASH) data repository (n = 431), Massachusetts General Hospital (MGH) (n = 91,237),	(UMLS) Metathesaurus, Semantic Network, and learning algorithms, frequency-inverse document frequency (Tf-idf)	Accuracy of 0.957 For (iDASH) data	2017
				Accuracy 0.964 For (MGH)	
4	Caccamisi et al. [25].	Data on patients' smoking status from EMRs. 85,000 classified sentences	Support Vector Machine Sequential Minimal Optimization (SMO)	Accuracy.98.14%	2020
				F-score 0.981	
5	Hammoud et al.[36].	2000 articles, with 10 classes diseases.in Arabic language	the Support Vector Machine (SVM), ABioNER model	ABioNER F1 97.433 Validation.	Russia 2021
				ABioNER F1 95.9124 Testing.	
				SVM was 89.1308 in Validation	
				SVM 87.3473 in Testing.	
6		5448 medical reports from EMR DATASET	LSVM ML-Perceptron Logistic regression With feature selection	99.39 Accuracy	Iraq 2021
				99.27 F1-Measure	
				99.16 precession	
				98.71 Recall	

In the third model, the results show that deep learning can achieve better and more accurate results than the classic algorithms, as shown in table (5.14), where the accuracy may reach 99%, but with a difference in the execution time. In this case, was much more than the rest of the algorithms and also needs more advanced processors. Applying all these layers to the data classified the medical reports into ten classes that resulted in high accuracy equal to (99.00%), F1-Measure (97.82%), precision (98.64%), and Recall (97.11%).

In general, promising results emerged for all applied algorithms. The best accuracy achieved was with the ML-Perceptron at (98.81%), this also offered a reduction in the amount of time required for training where feature selection was used.

# **CHAPTER 6**

## **CONCLUSION AND FUTURE WORK**

### **6.1 Conclusion**

Natural language processing is an effective way to improve the processing for any medical datasets. It helps to improve the accuracy and efficiency of any process performed on this data, for example extracting patient's information from EMR and determining a patient's medical classification and extract relevant data from electronic records and deciding about diagnosis of the disease and medication used, as this offers high reliability and accuracy, helping create better clinical databases.

A pre-processing step using NLP helps to make the initial unstructured data amenable to processing and mining and rids it of excess and useless attachments in order to allow greater benefit to be derived from the medical information within the initial records.

In the first model, diagnosing asthma through semi-structured data that contains a structured part and unstructured part. the proposed addition of TF-IDF has proven effective in identifying the best features of data and removing inappropriate or negative influences on unstructured data based on the use of NLP techniques. Overall, this offered promising results for all the algorithms examined.

Even a design method that shows great performance in general contexts may suffer from performance variation in specific biomedical fields. Applying all these layers to the data that classified the medical reports into ten classes have been resulted in high accuracy equal to (99%), F1-Measure (97.82%), precision (98.64%), and Recall (97.11%).

In the second model, was the classification of specialties in textual medical reports, a consecutive series of 3318 medical reports from the EMR dataset was evaluated. NLP with TF-IDF and Chi-Square FS techniques was implemented to train the algorithm to classify items into nine medical groups.

An FS step was added to reduce the number of features, retaining only the most effective features. The addition of this step significantly contributed to increase the accuracy of the implementation of all the algorithms used and reducing execution time. Various classification methods were applied to classify the dataset; the highest accuracy was achieved by applying Multi-Layer Perceptron classification techniques.

In the third model, the results show that deep learning can achieve better and more accurate results than the classic algorithms, Processing tools should be selected according to the characteristics of the data and the principles of dataset design followed.

## 6.2 Future work

- In this thesis, a classification system was established for patients that directs them to the department concerned with the disease, based on the text medical report. In the future, it is possible to work on expanding the work of the model to be able to direct patients through communication networks via the Internet, by converting the program's input into a text file that the application can deal with.
- In diagnosing asthma, this model classified the patients into two categories, asthmatic or not. In the future, trying to classify asthma based on the severity of the disease into four types: severe, moderate, mild, and non-affected, but this process requires obtaining a database larger than the currently available, which is the reason why the multiple classifications is not currently applied because an acceptable accuracy was not reached in the case of classifying the currently available data, which encourages the creation of a database for asthma that is larger than the current one.
- This thesis is the first of its kind in Iraq that discusses the use of NLP in analyzing data for Iraqi peoples and comparing the results with global data. The main reason is that health institutions in Iraq's lack of a reliable and comprehensive medical record. Therefore, it may be considered the nucleus for the creation of a medical and health record for patients in the future, based on correct foundations that can be processed electronically in a smooth manner that avoids the errors that occurred when creating unstructured and unclassified health records in other countries.



## References

- [1] W.-T. Wu *et al.*, “Data mining in clinical big data: the frequently used databases, steps, and methodological models,” *Mil. Med. Res.*, vol. 8, no. 1, pp. 1–12, 2021.
- [2] R. Bhardwaj, A. R. Nambiar, and D. Dutta, “A study of machine learning in healthcare,” in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 2017, vol. 2, pp. 236–241.
- [3] A. M. Nancy and R. Maheswari, “A Review On Unstructured Data In Medical Data,” *J. Crit. Rev.*, vol. 7, no. 13, pp. 2202–2208, 2020.
- [4] M. Tayefi *et al.*, “Challenges and opportunities beyond structured data in analysis of electronic health records,” *Wiley Interdiscip. Rev. Comput. Stat.*, p. e1549, 2021.
- [5] R. Attrey and A. Levit, “The promise of natural language processing in healthcare,” *Univ. West. Ont. Med. J.*, vol. 87, no. 2, pp. 21–23, 2018.
- [6] G. S. Alarcon *et al.*, “Risk factors for methotrexate-induced lung injury in patients with rheumatoid arthritis: a multicenter, case-control study,” *Ann. Intern. Med.*, vol. 127, no. 5, pp. 356–364, 1997.
- [7] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, “Health big data analytics: current perspectives, challenges and potential solutions,” *Int. J. Big Data Intell.*, vol. 1, no. 1–2, pp. 114–126, 2014.
- [8] R. S. H. Istepanian and T. Al-Anzi, “m-Health 2.0: new perspectives on mobile health, machine learning and big data analytics,” *Methods*, vol. 151, pp. 34–40, 2018.

- [9] T. Vos *et al.*, “Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019,” *Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.
- [10] S. Kukreja, “A Comprehensive Study on the Applications of Artificial Intelligence for the Medical Diagnosis and Prognosis of Asthma,” *Available SSRN 3081746*, 2017.
- [11] S. Mathur and B. Joshi, “Application of naïve bayes classification for disease prediction,” *Int. J. Manag. IT Eng.*, vol. 9, no. 4, pp. 80–87, 2019.
- [12] S. Kukreja, “A Comprehensive Study on the Applications of Machine Learning for the Medical Diagnosis and Prognosis of Asthma,” *arXiv Prepr. arXiv1804.04612*, 2018.
- [13] S. T. Wu *et al.*, “Automated chart review for asthma cohort identification using natural language processing: an exploratory study,” *Ann. Allergy, Asthma Immunol.*, vol. 111, no. 5, pp. 364–369, 2013.
- [14] Q. Do, T. C. Son, and J. Chaudri, “Classification of asthma severity and medication using TensorFlow and multilevel databases,” *Procedia Comput. Sci.*, vol. 113, pp. 344–351, 2017.
- [15] C.-I. Wi *et al.*, “Application of a natural language processing algorithm to asthma ascertainment. An automated chart review,” *Am. J. Respir. Crit. Care Med.*, vol. 196, no. 4, pp. 430–437, 2017.
- [16] S. K. Prabhakar and D.-O. Won, “Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention,” *Comput. Intell. Neurosci.*, vol. 2021, 2021.
- [17] A. K. Jha *et al.*, “Use of electronic health records in US hospitals,” *N. Engl. J. Med.*, vol. 360, no. 16, pp. 1628–1638, 2009.

- [18] S. Meystre and P. J. Haug, "Evaluation of medical problem extraction from electronic clinical documents using MetaMap Transfer (MMTx)," *Stud. Health Technol. Inform.*, vol. 116, pp. 823–828, 2005.
- [19] S. Meystre and P. J. Haug, "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation," *J. Biomed. Inform.*, vol. 39, no. 6, pp. 589–599, 2006.
- [20] F. R. Lucini *et al.*, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *Int. J. Med. Inform.*, vol. 100, pp. 1–8, 2017.
- [21] J. Xie, Y. Li, N. Wang, L. Xin, Y. Fang, and J. Liu, "Feature selection and syndrome classification for rheumatoid arthritis patients with Traditional Chinese Medicine treatment," *Eur. J. Integr. Med.*, vol. 34, p. 101059, 2020.
- [22] I. Li *et al.*, "Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review," *arXiv Prepr. arXiv2107.02975*, 2021.
- [23] C. Soguero-Ruiz *et al.*, "Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 5, pp. 1404–1415, 2014.
- [24] H. N. Alshaer, M. A. Otair, L. Abualigah, M. Alshinwan, and A. M. Khasawneh, "Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application," *Multimed. Tools Appl.*, vol. 80, no. 7, pp. 10373–10390, 2021.

- [25] A. Caccamisi, L. Jørgensen, H. Dalianis, and M. Rosenlund, “Natural language processing and machine learning to enable automatic extraction and classification of patients’ smoking status from electronic medical records,” *Ups. J. Med. Sci.*, vol. 125, no. 4, pp. 316–324, 2020.
- [26] J. de la Torre, J. Marin, S. Ilarri, and J. J. Marin, “Applying machine learning for healthcare: A case study on cervical pain assessment with motion capture,” *Appl. Sci.*, vol. 10, no. 17, p. 5942, 2020.
- [27] I. Solti, C. R. Cooke, F. Xia, and M. M. Wurfel, “Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches,” in *2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop*, 2009, pp. 314–319.
- [28] A. Esteva *et al.*, “A guide to deep learning in healthcare,” *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019.
- [29] S. Wu *et al.*, “Deep learning in clinical natural language processing: a methodical review,” *J. Am. Med. Informatics Assoc.*, vol. 27, no. 3, pp. 457–470, 2020.
- [30] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, “Medical text classification using convolutional neural networks,” in *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, IOS Press, 2017, pp. 246–250.
- [31] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, “Text Classification Based on Convolutional Neural Networks and Word Embedding for Low-Resource Languages: Tigrinya,” *Information*, vol. 12, no. 2, p. 52, 2021.

- [32] J. Geraci, P. Wilansky, V. de Luca, A. Roy, J. L. Kennedy, and J. Strauss, “Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression,” *Evid. Based. Ment. Health*, vol. 20, no. 3, pp. 83–87, 2017.
- [33] W.-H. Weng, K. B. Waghlikar, A. T. McCray, P. Szolovits, and H. C. Chueh, “Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach,” *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, pp. 1–13, 2017.
- [34] X. Mu and H. Zhang, “Embedded electronic medical record text data mining using neural network association classification algorithm,” *Expert Syst.*, p. e12874, 2021.
- [35] A. A. Thomas *et al.*, “Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results,” *World J. Urol.*, vol. 32, no. 1, pp. 99–103, 2014.
- [36] J. Hammoud, A. Vatian, N. Dobrenko, N. Vedernikov, A. Shalyto, and N. Gusarova, “New Arabic Medical Dataset for Diseases Classification,” *arXiv Prepr. arXiv2106.15236*, 2021.
- [37] M. Khachidze, M. Tsintsadze, and M. Archuadze, “Natural language processing based instrument for classification of free text medical records,” *Biomed Res. Int.*, vol. 2016, 2016.
- [38] R. Cohen, I. Aviram, M. Elhadad, and N. Elhadad, “Redundancy-aware topic modeling for patient record notes,” *PLoS One*, vol. 9, no. 2, p. e87555, 2014.
- [39] P. Garrett and J. Seidman, “EMR vs EHR—What is the Difference,” *Heal. January*, 2011.

- [40] C. Dobbing, “Paperless practice—electronic medical records at Island Health,” *Comput. Methods Programs Biomed.*, vol. 64, no. 3, pp. 197–199, 2001.
- [41] R. Dale, H. Moisl, and H. Somers, *Handbook of natural language processing*. CRC press, 2000.
- [42] B. K. Sidhu, “Natural language processing,” *Int. J. Comput. Technol. Appl.*, vol. 4, no. 5, p. 751, 2013.
- [43] K. Chowdhary, “Natural language processing,” *Fundam. Artif. Intell.*, pp. 603–649, 2020.
- [44] F. Sun, A. Belatreche, S. Coleman, T. M. McGinnity, and Y. Li, “Pre-processing online financial text for sentiment classification: A natural language processing approach,” in *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2014, pp. 122–129.
- [45] S. Kannan *et al.*, “Preprocessing techniques for text mining,” *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2014.
- [46] T. Ganegedara, *Natural Language Processing with TensorFlow: Teach language to machines using Python’s deep learning library*. Packt Publishing Ltd, 2018.
- [47] D. Munková, M. Munk, and M. Vozár, “Data pre-processing evaluation for text mining: transaction/sequence model,” *Procedia Comput. Sci.*, vol. 18, pp. 1198–1207, 2013.
- [48] A. Aichert, “Feature extraction techniques,” in *Camp medical seminar ws0708*, 2008, pp. 1–8.

- [49] K. Soumya George and S. Joseph, “Text classification by augmenting bag of words (BOW) representation with co-occurrence feature,” *IOSR J. Comput. Eng.*, vol. 16, no. 1, pp. 34–38, 2014.
- [50] W. Scott, “TF-IDF from scratch in python on real world dataset.” 2020.
- [51] “[Preprocessing techniques for text mining-an overview,” Int.” .
- [52] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [53] J. Brownlee, “How to choose a feature selection method for machine learning,” *Mach. Learn. Mastery*, vol. 10, 2019.
- [54] N. AlNuaimi, M. M. Masud, M. A. Serhani, and N. Zaki, “Streaming feature selection algorithms for big data: A survey,” *Appl. Comput. Informatics*, 2020.
- [55] J. Fan, R. Samworth, and Y. Wu, “Ultrahigh dimensional feature selection: beyond the linear model,” *J. Mach. Learn. Res.*, vol. 10, pp. 2013–2038, 2009.
- [56] W. Jia, M. Sun, J. Lian, and S. Hou, “Feature dimensionality reduction: a review,” *Complex Intell. Syst.*, pp. 1–31, 2022.
- [57] I. S. Thaseen and C. A. Kumar, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM,” *J. King Saud Univ. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017.
- [58] A. J. Ferreira and M. A. T. Figueiredo, “Efficient feature selection filters for high-dimensional data,” *Pattern Recognit. Lett.*, vol. 33, no. 13, pp. 1794–1804, 2012.

- [59] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, 2016.
- [60] P. Sharma, “The ultimate guide to 12 dimensionality reduction techniques (with Python codes),” *Anal. Vidhya*, p. 27, 2018.
- [61] M. Scholz, “Approaches to analyse and interpret biological profile data.” Universität Potsdam, 2006.
- [62] E. Alpaydin, *Machine learning: the new AI*. MIT press, 2016.
- [63] J. Stuart, “Artificial Intelligence A Modern Approach Third Edition.” Prentice Hall, 2010.
- [64] M. Natarajan, “Role of text mining in information extraction and information management,” *DESIDOC J. Libr. Inf. Technol.*, vol. 25, no. 4, 2005.
- [65] S. J. Lee and K. Siau, “A review of data mining techniques,” *Ind. Manag. Data Syst.*, 2001.
- [66] V. Kotu and B. Deshpande, “Recommendation engines,” *data Sci. (Second Ed.*, pp. 343–394, 2019.
- [67] D. Milward, “What Is Text Mining Text Analytics and Natural Language Processing?,” *Linguamatics*, 2020.
- [68] H. M. Habeeb, “An overview on the use of data mining and linguistics techniques for building microblog-based early detection systems in the healthcare sector,” *AIRCC’s Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 5, pp. 143–155, 2015.



- [69] M. Awad and R. Khanna, “Support vector machines for classification,” in *Efficient Learning Machines*, Springer, 2015, pp. 39–66.
- [70] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, 2020.
- [71] T. S. R. Sai, “Comparing The Performance Of Various SVM Classification Techniques: A Survey,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 13, pp. 1129–1136, 2021.
- [72] G. Xiao, J. Xing, and Y. Zhang, “Surface roughness prediction model of GH4169 superalloy abrasive belt grinding based on multilayer perceptron (MLP),” *Procedia Manuf.*, vol. 54, pp. 269–273, 2021.
- [73] K.-L. Du and M. N. S. Swamy, “Multilayer perceptrons: Architecture and error backpropagation,” in *Neural Networks and Statistical Learning*, Springer, 2014, pp. 83–126.
- [74] A. Mohanty, “Multi layer Perceptron (MLP) Models on Real World Banking Data.” Medium. [https://becominghuman.ai/multi-layer-perceptronmlp-models-on-real ...](https://becominghuman.ai/multi-layer-perceptronmlp-models-on-real-...), 2019.
- [75] S. Agarwal, G. N. Pandey, and M. D. Tiwari, “Data mining in education: data classification and decision tree approach,” *Int. J. e-Education, e-Business, e-Management e-Learning*, vol. 2, no. 2, p. 140, 2012.
- [76] X. Wu and V. Kumar, *The top ten algorithms in data mining*. CRC press, 2009.

- [77] A. Priyam, G. R. Abhijeeta, A. Rathee, and S. Srivastava, “Comparative analysis of decision tree classification algorithms,” *Int. J. Curr. Eng. Technol.*, vol. 3, no. 2, pp. 334–337, 2013.
- [78] Y. Tedla and K. Yamamoto, “Analyzing word embeddings and improving POS tagger of tigrinya,” in *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 115–118.
- [79] J. A. Sparks *et al.*, “Rheumatoid arthritis disease activity predicting incident clinically apparent rheumatoid arthritis–associated interstitial lung disease: a prospective cohort study,” *Arthritis Rheumatol.*, vol. 71, no. 9, pp. 1472–1482, 2019.
- [80] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep Learning--based Text Classification: A Comprehensive Review,” *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2021.
- [81] R. J. Roberts, “PubMed Central: The GenBank of the published literature,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 2. National Acad Sciences, pp. 381–382, 2001.
- [82] J.-D. Kim, T. Ohta, and J. Tsujii, “Corpus annotation for mining biomedical events from literature,” *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–25, 2008.
- [83] S. Christin, É. Hervet, and N. Lecomte, “Going further with model verification and deep learning,” *Methods Ecol. Evol.*, vol. 12, no. 1, pp. 130–134, 2021.
- [84] W.-C. Kang *et al.*, “Learning to embed categorical features without embedding tables for recommendation,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 840–850.

- [85] S. S. S. KS K, “A Survey of Embeddings in Clinical Natural Language Processing,” *arXiv Prepr. arXiv1903.01039*, 2019.
- [86] G. Li, X. Du, X. Li, L. Zou, G. Zhang, and Z. Wu, “Prediction of DNA binding proteins using local features and long-term dependencies with primary sequences based on deep learning,” *PeerJ*, vol. 9, p. e11262, 2021.
- [87] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [88] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Information*, vol. 10, no. 4, p. 150, 2019.
- [89] E. Voita, R. Sennrich, and I. Titov, “Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT,” *arXiv Prepr. arXiv2109.01396*, 2021.
- [90] H. Gholamalinezhad and H. Khosravi, “Pooling methods in deep neural networks, a review,” *arXiv Prepr. arXiv2009.07485*, 2020.
- [91] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
- [92] M. Hassler and G. Fliedl, “Text preparation through extended tokenization,” *WIT Trans. Inf. Commun. Technol.*, vol. 37, 2006.
- [93] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

- [94] J. Jeong, “The most intuitive and easiest guide for convolutional neural network,” Available in: <https://towardsdatascience.com/the-most-intuitive-and-easiest-guide-for-convolutional-neural-network-3607be47480>. Cited, 2019.
- [95] R. Ghosh and A. K. Gupta, “Scale steerable filters for locally scale-invariant convolutional neural networks,” *arXiv Prepr. arXiv1906.03861*, 2019.
- [96] N. Petkov, “Automatic segmentation of indoor and outdoor scenes from visual lifelogging,” in *Applications of Intelligent Systems: Proceedings of the 1st International APPIS Conference 2018*, 2018, vol. 310, p. 194.
- [97] M. Niu, Y. Li, C. Wang, and K. Han, “RFAmyloid: a web server for predicting amyloid proteins,” *Int. J. Mol. Sci.*, vol. 19, no. 7, p. 2071, 2018.
- [98] N. Pearce *et al.*, “Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC),” *Thorax*, vol. 62, no. 9, pp. 758–766, 2007.
- [99] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.
- [100] E. Loper and S. Bird, “Nltk: The natural language toolkit,” *arXiv Prepr. cs/0205028*, 2002.
- [101] A. M. Khan and K. R. Afreen, “An approach to text analytics and text mining in multilingual natural language processing,” *Mater. Today Proc.*, 2021.
- [102] G. Nunberg, *The linguistics of punctuation*, no. 18. Center for the Study of Language (CSLI), 1990.

## الخلاصة

تعد معالجة اللغة الطبيعية جزءًا من خوارزميات الذكاء الاصطناعي التي تركز على تصميم وبناء التطبيقات والأنظمة بطريقة تسمح بالتفاعل بين أجهزة الكمبيوتر واللغات الطبيعية المطورة للاستخدام البشري، وقد تم استخدام البرمجة اللغوية العصبية في عدة مجالات ضمن الذكاء الاصطناعي ومعالجة البيانات والتطبيقات مثل تطبيقات الوسائط الاجتماعية والتطبيقات الطبية وتطبيقات الترجمة مما كان له أثر إيجابي في تحسين جودة البيانات واستخراج المعلومات المفيدة في معظم التطبيقات.

تم اقتراح نماذج للاستفادة من البيانات الطبية المتوفرة على شكل ملفات نصية في تشخيص مرض الربو، الربو هو مرض التهابي مزمن شائع جدا في المجتمع ينتج عنه تضيق في الشعب الهوائية مع تأثير كبير على الأطفال والبالغين بسبب ارتفاع معدلات الاعتلال والوفيات في الحالات الشديدة.

في هذا النموذج، سيتم استخدام قاعدة بيانات شبه منظمة للمرضى الصغار. يتكون النموذج المقترح من أربع مراحل رئيسية. الأول هو جمع البيانات والتحضير لعملية التعدين. والثاني هو المعالجة المسبقة للبيانات والتي تم إجراؤها من خلال تطبيق خوارزميات مختلفة لمعالجة اللغة الطبيعية وتتضمن المرحلة الثالثة استخراج الميزات وترجيحها من خلال تطبيق الأداة (TF-IDF).

قمنا بتحويل الجزء غير المهيكل من البيانات إلى مهيكل بواسطة أدوات البرمجة اللغوية العصبية. ثم تطبيق خوارزميات التصنيف عليها. يتم إدخال الميزات المستخرجة في تقنيات التعلم الآلي للتشخيص كمرحلة نهائية. أظهرت النتائج تحقيق دقة عالية بعد تطبيق خوارزميات معالجة اللغة الطبيعية حيث كانت أعلى دقة تم التوصل إليها في خوارزمية (ML-Perceptron) (٩٩,٨٩٪) و (٩٧,٥١٪) بتطبيق على مجموعة بيانات غرينادا ومجموعة البيانات العراقية على التوالي.

أما النموذج الثاني هو تصنيف التخصصات في التقارير الطبية النصية، واستخدمت طرق استخراج الميزات واختيار الميزات لتحويل التقارير الطبية النصية إلى مجموعات من الميزات واستخراج الميزات الأكثر فعالية. تم تطبيق طرق تصنيف مختلفة لتصنيف مجموعة البيانات؛ تم تحقيق أعلى دقة من خلال تطبيق خوارزمية (ML-Perceptron) حيث بلغت (٩٩,٣٩) %.

يطبق النموذج الأخير خوارزمية التعلم العميق Convolution Neural Network (CNN) على نفس مجموعة بيانات التقرير الطبي النصي المستخدمة في النموذج السابق. تطبيق البرمجة اللغوية العصبية لتنظيف البيانات والشبكة العصبية التي تتكون من خمس طبقات. بتطبيق كل هذه الطبقات على بياناتنا، تمكن النموذج من تصنيف التقارير الطبية إلى عشر فئات نتج عنها دقة عالية تساوي (99%)



جمهورية العراق  
وزارة التعليم العالي والبحث العلمي  
جامعة كربلاء/كلية الهندسة  
قسم الهندسة الكهربائية والإلكترونية

## معالجة اللغة الطبيعية وتعلم الآلة في تحليل التقارير الطبية

رسالة مقدمه الى

قسم الهندسة الكهربائية والإلكترونية في كلية الهندسة / جامعة كربلاء

كجزء من متطلبات نيل شهادة الماجستير في علوم الهندسة الكهربائية والإلكترونية

من قبل الطالب

حسنين عبد الجواد حسين علي المحنه

بكالوريوس هندسة كهربائية / كلية الهندسة / جامعة بابل

تحت إشراف

أ. د حوراء حسن عباس

محرم ١٤٤٤

آب ٢٠٢٢