

ANALYSIS AND PREDICTION OF CHRONIC KIDNEY DISEASE

Thilina Duminda Nakkawita

(168249N)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2020

ANALYSIS AND PREDICTION OF CHRONIC KIDNEY DISEASE

Thilina Duminda Nakkawita

(168249N)

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree
Master of Science

Department of Computer Science and Engineering

University of Moratuwa


Sri Lanka

April 2020

DECLARATION

I declare that the thesis is purely based on my own work, and it does not include course materials from any other university or college diploma without acknowledgement, and as per my knowledge this does not include Materials that are published or written by other persons, unless otherwise not indicated in the text.

In addition, I give the non-exclusive right to the University of Moratuwa to replicate or distribute my article in whole or in part in print, electronic or other media. I reserve the right to use this content in whole or part in future works.



2020-05-30

.....
Signature

.....
Date

I confirm that, as per my knowledge, the above statement of the candidate is correct and that the project report can be used to evaluate the MSc research project.

.....
Signature of the supervisor

.....
Date

ABSTRACT

In Sri Lanka, chronic kidney disease has become a significant public health problem over the past two decades. Since there are few signs or symptoms in the early stages, it is difficult to identify whether people have the CKD disease, because. Due to this reason, they do not get treatments. If the disease is detected at an early stage, CKD can be cured. Sri Lanka currently lacks comprehensive and systematic surveillance procedures to identify and monitor all aspects of CKD in the general population.

The disease can be identified in the early stages if there is a proper dataset to analyze. Based on the data a predictive model can be developed and this will help doctors diagnose if a patient has early-stage CKD. CKD can prevent if this detects early and provide necessary treatments.

As part of my research; I have developed a computerized system to capture and track aspects of CKD in Sri Lanka, including a predictive model to detect CKD in its early stages. The predictive model was developed using different types of data mining classification algorithms. In the healthcare sector, data mining is mainly used for disease detection. Broad data mining techniques exist for predicting diseases, such as classification, clustering, association rules, summaries, and regression. Additionally, the tool was developed to perform several analyses based on the collected data.

ACKNOWLEDGEMENTS

I would like to take this opportunity to share my sincere gratitude for the people who have been instrumental in the completion of this dissertation.

To Dr. Shehan Perera, I cannot say Thank You enough for all of your support, guidance, and encouragement: you have motivated and enlightened me more than you know throughout this project. It would not have been possible without your help.

To all of the other individuals who have helped and contributed on this project, I Thank You for your incredible assistance and involvement on this dissertation.

TABLE OF CONTENTS

DECLARATION.....	i
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vi
LIST OF TABLE	viii
LIST OF ABBREVIATIONS	ix
1. INTRODUCTION.....	1
1.1 Chronic Kidney Disease (CKD)	1
1.2 Chronic Kidney Disease (CKD) in Sri Lanka.....	3
1.3 Risk Factors of CKDu	3
1.4 Prediction by Data Mining.....	4
2. RESEARCH PROBLEM & OBJECTIVES.....	5
2.1 Centralized System to Collect Data	6
2.2 Analysis of Patient Data	6
2.2.1. Analysis of patient data according to AGA Divisions.....	7
2.3 Building a Predictive Model.....	10
2.4 Summary.....	11
3. LITERATURE REVIEW.....	12
3.1 About the CKD.....	12
3.2 Analytics.....	13
3.2.1 Descriptive Analysis	13
3.2.2 Predictive Analysis	16
3.3 Summary.....	20
4. RESEARCH MODEL / METHODOLOGY	21
4.1 Build a System to Collect Data.....	21
4.2 Dataset.....	23
4.3 Analyze the Collected Data	23
4.4 Prediction Based on Analysis.....	28
4.5 Classification Techniques.....	30
4.6 Best Algorithm	30
4.6.1 J48 Algorithm.....	31
4.6.2 Zero Algorithm	31

4.6.3 Naive Bayes Algorithm	32
4.6.4 Hoeffding Tree Algorithm	33
4.6.5 Decision Table Algorithm	34
4.6.6 Random Forest Algorithm	35
4.7 Comparison of the Results	36
4.8 Summary.....	37
5. SYSTEM/SOLUTION ARCHITECTURE AND IMPLEMENTATION.....	38
5.1 Patient Database.....	39
5.2 Prediction Tool.....	39
5.3 Patient Data Analysis.....	40
5.4 Summary.....	41
6. SYSTEM EVALUATION (DATA AND ANALYSIS)	42
6.1 Data Pre-processing.....	42
6.2 Results / Outcome of the Prediction Tool	45
6.3 Validate the Results of the Prediction Tool	45
6.4 Patient Data Analysis Tool.....	48
6.5 Summary.....	49
7. CONCLUSION	50
7.1 Future Works	53
REFERENCES.....	54

LIST OF FIGURES

Figure 2.1	Centralized system	7
Figure 2.2	Patient data according to AGA Divisions	8
Figure 2.3	Impact of associated medical conditions of the patient	9
Figure 2.4	Patient data according to the lab data	10
Figure 2.5	Prediction with dataset	11
Figure 4.1	System	26
Figure 4.2	Patient entry form – Basic Information	27
Figure 4.3	Patient entry form - Diagnosis data entry form	27
Figure 4.4	Analyze patient data according to AGA	29
Figure 4.5	Analyze patient data according to Medical Condition	30
Figure 4.6	Analyze patient data according to Proteinuria	31
Figure 4.7	Analyze patient data according to Serum albumin	31
Figure 4.8	Analyze patient data according to PTH Level	32
Figure 4.9	Analyze patient data according to the age range	33
Figure 4.10	Prediction tool results	34
Figure 4.11	XML Response	35
Figure 4.12	Results using J48 Algorithm	36
Figure 4.13	Results using ZeroR Algorithm	37
Figure 4.14	Results using Naïve Bayes Algorithm	38
Figure 4.15	Results using Hoeffding Tree Algorithm	39
Figure 4.16	Results using the Decision Table Algorithm	40
Figure 4.17	Results using the Random Forest Algorithm	41
Figure 5.1	System architecture	43
Figure 5.2	Sample WEKA code	45
Figure 5.3	WEKA results	46
Figure 5.4	BI sample code	46
Figure 6.0	Training Dataset	48
Figure 6.1	Correlation based feature selection	49
Figure 6.2	Accuracy of model	49

Figure 6.3	Screenshot of Prediction Tool	50
Figure 6.4	Prediction Results	
Figure 6.5	Weka Results for CKD Patients	51
Figure 6.6	Weka Results for non-CKD Patients	52
Figure 6.7	Analysis Tool	53
Figure 6.8	Power BI Chart	54

LIST OF TABLE

		Page
Table 4.1	Summary of WEKA results	42
Table 5.1	Attributes of CKD Patients	44
Table 6.1	Selected attributes for predictive data model	47
Table 7.1	Attributes of patient record used by prediction model	55
Table 7.2	Summary of WEKA results	56

LIST OF ABBREVIATIONS

AGA	Assistant Government Agents
API	Application Programming Interface
CKD	Chronic Kidney Disease
CKDu	Chronic Kidney Disease Unknown
CRF	Chronic Renal Failure
eGFR	Estimated Glomerular Filtration Rate
GFR	Glomerular Filtration Rate
PTH	Para Thyroid Hormone