MULTI-LABEL RISK DIABETES COMPLICATION PREDICTION MODEL USING DEEP NEURAL NETWORK WITH MULTI-CHANNEL WEIGHTED DROPOUT

NUR RACHMAN DZAKIYULLAH

DOCTOR OF PHILOSOPHY

2025

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Information and Communication Technology

MULTI-LABEL RISK DIABETES COMPLICATION PREDICTION MODEL USING DEEP NEURAL NETWORK WITH MULTI-CHANNEL WEIGHTED DROPOUT

Nur Rachman Dzakiyullah

Doctor of Philosophy

2025

MULTI-LABEL RISK DIABETES COMPLICATION PREDICTION MODEL USING DEEP NEURAL NETWORK WITH MULTI-CHANNEL WEIGHTED DROPOUT

NUR RACHMAN DZAKIYULLAH

**A thesis submitted
in fulfillment of the requirements for the degree of
Doctor of Philosophy**

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2025

DECLARATION

I declare that this thesis entitled “Multi-Label Risk Diabetes Complication Prediction Model Using Deep Neural Network with Multi-Channel Weighted Dropout” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:
Name	: NUR RACHMAN DZAKIYULLAH
Date	: 14 FEBRUARY 2025

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Doctor of Philosophy.

Signature	:
Supervisor Name	: PROFESSOR TS. DR. BURHANUDDIN BIN MOHD ABOOBAIDER
Date	: 14 FEBRUARY 2025

DEDICATION

“To my beloved Mother and Father”

“To my beloved Mother and Father-in-Law”

” To my beloved wife”

“To my beloved sons”

“To my beloved brother and sister”

ABSTRACT

The early diagnosis of diabetes complications using risk factors remains underexplored, particularly with the application of Multi-Label Classification (MLC). This study addresses this gap by leveraging data from the Behavioral Risk Factor Surveillance System (BRFSS) from 2016 to 2021 to categorize seven diabetes complications simultaneously. By employing Artificial Intelligence (AI), this study examines the interconnected nature of these complications. A total of 33 variables from each year of the BRFSS dataset were analyzed, incorporating statistical techniques to understand the data and preprocessing methods to prepare it for machine learning. Seven machine learning models—Artificial Neural Network (ANN), Random Forest (RF), Decision Tree (DTT), k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Support Vector Machine (SVM), and Deep Neural Network (DNN)—were used for multi-label classification of the complications. The study employed two MLC frameworks: Problem Transformation methods (Binary Relevance, Classifier Chains, Label Power Set, and Calibrated Label Ranking) and Algorithm Adaptation. Performance was evaluated using 20 metrics, including AUROC and other advanced indicators. The first experiment revealed that the Algorithm Adaptation framework outperformed Problem Transformation methods across most metrics. Among the models, the DNN achieved superior performance in key metrics such as Subset Accuracy (0.4156), Hamming Loss (0.1272), F1-Score (macro) (0.9113), and AUROC (macro) (0.7935). Feature importance analysis identified the top 10 variables influencing different complications. The second experiment introduced a novel dropout regularization technique called multi-channel weighted dropout, designed to enhance model generalization. Comparative evaluations with existing dropout methods demonstrated the superior performance of the proposed technique, particularly when applied within the Algorithm Adaptation framework using DNNs. The proposed method managed data and model complexity effectively while maintaining high computational efficiency. This study contributes to the field by proposing a new regularization technique, demonstrating the effectiveness of the Algorithm Adaptation framework, and providing insights into the associations between diabetes complications. These findings highlight the potential of AI-driven MLC approaches in advancing diabetes complication.

MODEL RAMALAN KOMPLIKASI DIABETES RISIKO MULTILABEL MENGGUNAKAN RANGKAIAN NEURAL DALAM DENGAN KECICIRAN PEMBERAT MULTISALURAN

ABSTRAK

Diagnosis awal komplikasi diabetes berdasarkan faktor risiko masih kurang diterokai, terutamanya dengan penggunaan Multi-Label Classification (MLC). Kajian ini menangani jurang tersebut dengan memanfaatkan data daripada Behavioral Risk Factor Surveillance System (BRFSS) bagi tahun 2016 hingga 2021 untuk mengkategorikan tujuh komplikasi diabetes secara serentak. Dengan menggunakan kuasa pengkomputeran Artificial Intelligence (AI), kajian ini mengkaji hubungan antara komplikasi tersebut. Sebanyak 33 pemboleh ubah dari setiap tahun data BRFSS dianalisis, menggunakan teknik statistik untuk memahami data dan kaedah prapemprosesan untuk menyediakan data bagi pembelajaran mesin. Tujuh model pembelajaran mesin—Artificial Neural Network (ANN), Random Forest (RF), Decision Tree (DTT), k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Support Vector Machine (SVM), dan Deep Neural Network (DNN)—digunakan untuk klasifikasi pelbagai label komplikasi. Kajian ini menggunakan dua kerangka MLC: kaedah Transformasi Masalah (Binary Relevance, Classifier Chains, Label Power Set, dan Calibrated Label Ranking) dan Adaptasi Algoritma. Prestasi dinilai menggunakan 20 metrik, termasuk AUROC dan indikator lanjutan lain. Eksperimen pertama menunjukkan bahawa kerangka Adaptasi Algoritma mengatasi kaedah Transformasi Masalah dalam kebanyakan metrik. Dalam kalangan model, DNN mencatatkan prestasi terbaik dalam metrik utama seperti Subset Accuracy (0.4156), Hamming Loss (0.1272), F1-Score (macro) (0.9113), dan AUROC (macro) (0.7935). Analisis kepentingan ciri mengenal pasti 10 pemboleh ubah utama yang mempengaruhi pelbagai komplikasi. Eksperimen kedua memperkenalkan teknik regularisasi dropout baharu yang dinamakan multi-channel weighted dropout, yang direka untuk meningkatkan generalisasi model. Penilaian perbandingan dengan kaedah dropout sedia ada menunjukkan prestasi unggul teknik yang dicadangkan, terutamanya apabila digunakan dalam kerangka Adaptasi Algoritma dengan DNN. Kaedah yang dicadangkan berjaya menguruskan kerumitan data dan model dengan berkesan sambil mengekalkan kecekapan pengiraan yang tinggi. Kajian ini menyumbang kepada bidang ini dengan mencadangkan teknik regularisasi baharu, membuktikan keberkesanan kerangka Adaptasi Algoritma, dan memberikan wawasan tentang hubungan antara komplikasi diabetes. Penemuan ini menonjolkan potensi pendekatan MLC yang didorong oleh AI dalam meningkatkan pengurusan diabetes.

ACKNOWLEDGEMENT

I express my profound gratitude to Allah the Almighty for His innumerable blessings.

I wish to extend my deepest appreciation to my primary supervisor, Prof. Dr. Burhanuddin Mohd. Aboobaider, and my secondary supervisor, Dr. Raja Rina Binti Raja Ikram, for their indispensable guidance and insightful counsel throughout my dissertation journey.

I sincerely thank all professors, staff, and technicians of the Faculty of Information and Communication Technology (FTMK-UTeM) for their unwavering support in helping me navigate technical challenges during my studies. I am also deeply grateful to my fellow Indonesian and Malaysian students, particularly the OPTIMASS and BIOCORE Members, for their guidance, encouragement, camaraderie, and both material and spiritual support during the preparation of my thesis.

I am profoundly appreciative of my colleagues at Universitas Alma Ata, especially those from the Faculty of Computer and Engineering (FKT-UAA), for their encouragement, collaboration, and invaluable support throughout this journey. Special recognition goes to the Rector of Universitas Alma Ata, Prof. Dr. H. Hamam Hadi, MS., Sc.D., Sp.GK., for his steadfast support and inspiration.

I will forever be indebted to my beloved parents, the late alm. Prof. Dr. Ir. R. Chairul Saleh, M.Sc., Ph.D., and Raden Ajeng Diah Purnamawati, for their unwavering prayers, love, and patience. My heartfelt thanks also go to my brother and his wife, Raden Achmad Chairdino Leuveano, S.T., M.Sc., Ph.D., and Tatbita Titin Suhariyanto, S.T., M.Sc., as well as my beloved sister and her husband, Raden Ajeng Dinovita Nurulhaq, S.I.Kom., M.Sc., and Muhammad Nur Alfani, S.Kom., for their steadfast support.

I extend my deepest appreciation to my beloved wife, Putri Rahmasari, S.ST., MPH, who has always stood by my side with unwavering faith, understanding, and patience. Finally, to my wonderful children, Muhammad Rosyid Alzam, Muhammad Fattah Rachman Saleh, and Buna Tsalisa Ahmad, who continue to inspire me and fill my life with positivity—I owe a significant part of this journey to them.

May Allah SWT bestow His blessings upon us all and reunite us in Paradise. Aamiin, Ya Allah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	-	Algorithm Adaptation
ADA	-	American Diabetes Association
AI	-	Artificial Intelligence
ANN	-	Artificial Neural Network
AUC	-	Area Under Curve
BR	-	Binary Relevance
BRFSS	-	Behavioural Risk Factor Surveillance System
CART	-	Classification And Regression Tree
CC	-	Classifier Chain
CDC	-	Centres For Disease Control and Prevention
CGM	-	Continuous Glucose Monitoring
CLR	-	Calibrated Label Ranking
CNN	-	Convolutional Neural Network
CRISP - DM	-	Cross Industry Standard Process for Data Mining
CVD	-	Cardiovascular Diseases
DCD	-	Diabetes Complications Dataset
DL	-	Deep Learning
DM	-	Diabetes Mellitus
DMT	-	Data Mining Techniques
DNN	-	Deep Neural Network
DT	-	Dropout Technique
DTT	-	Decision Tree Technique
EBM	-	Energy-Based Model
ECG	-	Electrocardiogram
ECG	-	Ensemble Classifier Chains
EDA	-	Exploratory Data Analysis
EHR	-	Electronic Health Records
FI	-	Feature Importance
FPG	-	Fasting Plasma Glucose
G - MEAN	-	Geometric Mean
GAN	-	Generative Adversarial Networks
GDM	-	Gestational Diabetes Mellitus
GWAS	-	Genome-Wide Association Studies
HR	-	Heart Rate
HRV	-	Heart Rate Variability
HSS	-	Heidke Skill Score
IDF	-	International Diabetes Federation

IFG	-	Impaired Fasting Glucose
IGT	-	Impaired Glucose Tolerance
k - NN	-	K-Neirest Neighbour
L1	-	Regularization Lasso Regression
L2	-	Ridge Regression
LP	-	Label Powerset
LR	-	Logistic Regression
LSTM	-	Long Short-Term Memory
MASHAD	-	Mashhad Stroke and Heart Atherosclerotic Disorders
MCC	-	Mathew's Correlation Coefficient
MCWD	-	Multi-Channel Weighted Dropout
ML	-	Machine Learning
ML - ANN	-	Multi-Label Artificial Neural Network
ML - DNN	-	Multi-Label Deep Neural Network
ML - DTT	-	Multi-Label Decision Tree
ML - kNN	-	Multi-Label K-Nearest Neighbours
ML - NB	-	Multi-Label Naïve Bayes
ML - RF	-	Multi-Label Random Forest
ML - SVP	-	Multi-Label Support Vector Machine
MLC	-	Multi-Label Classification
MLC - AA	-	Multi-Label Classification-Algorithm Adaptation
MLC - PT	-	Multi-Label Classification-Problem Transformation
MLP	-	Multi-Layer Perceptron
MLRDCP	-	Multi-Label Risk Diabetes Complication Prediction
MRI	-	Magnetic Resonance Imagery
MSE	-	Mean Square Error
NaN	-	Missing Values
NB	-	Naïve Bayes
NCDs	-	Non-Communicable Diseases
NLP	-	Natural Language Processing
OGTT	-	Oral Glucose Tolerance Test
PG	-	Plasma Glucose
PIDD	-	Pima Indians Diabetes Data
PPG	-	Photoplethysmogram
PSS	-	Peirce Skill Score
QUEST	-	The Quick Unbiased Efficient Statistical Tree
RDCP	-	Risk Diabetes Complication Prediction
RF	-	Random Forrest
RO	-	Research Objective
ROC	-	Receiver Operating Characteristic Curve
SHAP	-	Shapley Additive Explanations

SMOTE	-	Synthetic Minority Oversampling Technique
SNP	-	Single Nucleotide Polymorphisms
SOM	-	Self-Organizing Map
SVM	-	Support Vector Machine
T1DM	-	Type 1 Diabetes Mellitus
T2DM	-	Type 2 Diabetes Mellitus
TLGS	-	Tehran Lipid and Glucose Study
WHO	-	World Health Organization

LIST OF SYMBOLS

χ	-	d-dimensional instance space $\mathbb{R}^d$ $(or\ \mathbb{Z}^d)$
$\mathcal{Y}$	-	Label space with q possible class labels $\mathcal{Y} = \{y_1, y_2, \dots, y_q\}$
x	-	d-dimensional feature vector $(x_1, x_2, x_3, \dots . x_d)^{\mathrm{T}}, (x \in \chi)$
Y	-	Label set associated with x $(\subseteq \mathcal{Y})$
$\bar{Y}$	-	Complementary set of Y in $\mathcal{Y}$
$\mathcal{D}$	-	Multi-label training set $\{(x_i, Y_i) \mid 1 \leq i \leq m\}$
$\mathcal{S}$	-	Multi-label test set $\{(x_i, Y_i) \mid 1 \leq i \leq p\}$
$h(\cdot)$	-	Multi-label classifier $h: \chi \to 2^{\mathcal{Y}}$, where $h(x)$ return the set of proper labels for x
$f(\cdot,\cdot)$	-	Real value $f: \chi \times \mathcal{Y} \to \mathbb{R}, where\ f(x, y)$ return the confidence of y being proper label of x
$rank_f(\cdot,\cdot)$	-	$rank_f(x, y)$ return the rank of y in $\mathcal{Y}$ based on the decending order induced from $f(x,\cdot)$
$t(.)$	-	Thresholding function $t: \chi \to \mathbb{R}, where\ h(x) = \{y \mid f(x, y) > t(x), y \in \mathcal{Y}$
$\lvert . \rvert$	-	$\lvert A \rvert$ $return$ the cardinality of set A
$[\![.]\!]$	-	$[\![\pi]\!]$ return 1 if predicate π hold and 0 otherwise
$\emptyset(.,.)$	-	$\emptyset(Y, y)$return +1 if y $\in Y$ and -1 otherwise
$\mathcal{D}_j$	-	Binary training set $\{(x_i, \psi(Y_i, y_j, y_k) \mid \emptyset(Y_i, y_j) \neq \emptyset(Y_i, y_k), 1 \leq i \leq m)\}$ derived from $\mathcal{D}$ for label pair (y_j, y_k)
$\sigma_y(.)$	-	Injective function $\sigma_y: 2^{\mathcal{Y}} \to \mathbb{N}$ mapping from the power set of $\mathcal{Y}$ to natural number (σ_y^{-1} being the corresponding inverse function)
$\mathcal{D}_y^+$	-	Multi-class (single label) training set $\{(x_i, \sigma_y(Y_i)) \mid 1 \leq i \leq m)$ derived from $\mathcal{D}$
$\mathcal{B}$	-	Binary learning algorithm [complexity: $\mathcal{F}_{\mathcal{B}}(m, d)$ for training; $\mathcal{F}'_{\mathcal{B}}(d)$ for (per-instance) testing]
$\mathcal{M}$	-	Multi-class learning algorithm [complexity: $\mathcal{F}_{\mathcal{M}}(m, d, q)$ for training; $\mathcal{F}'_{\mathcal{M}}(d, q)$ for (per-instance) testing]
L	-	Hidden layer
$W^{(l)}$	-	The weight matrix
$b^{(l)}$	-	Bias vectors of the lth layer
y^l	-	The input of the lth layer
y^0	-	Network's input
W	-	Parameters learner
m	-	Dropout mask
M	-	Weighted dropout mask
Δ	-	Symmetric difference between sets
TP_j	-	Represent true positives for label j
FP_j	-	Represent false positives for label j

TN_j	-	Represent true negatives for label j
FN_j	-	Represent false negatives for label j
X_{input}	-	Represent Input number of dataset for BRFSS dataset
Y_{output}	-	Represent Output number of dataset for BRFSS dataset
n	-	the number of models
$Metric_i$	-	Represent total metric values for the i^{th}model.
λ_j		Class label for the j^{th}

LIST OF PUBLICATIONS

The followings are the list of publications related to the work on this thesis:

Dzakiyullah, N.R., Burhanuddin, M.A., Ikram, R.R., Ghani, K.A. and Setyonugroho, W., 2019. Machine learning methods for diabetes prediction. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), pp.2199-2205.

Dzakiyullah, N.R., Burhanuddin, M.A., Raja Ikram, R.R., Yudistira, N., Fauzi, M.R. and Purbohadi, D.J., 2024. Multi-Label Risk Prediction Diabetes Complication Using Machine Learning Models. *International Journal of Online & Biomedical Engineering*, 20(16).

CHAPTER 1

INTRODUCTION

1.1 Background

Artificial Intelligence (AI) possesses a robust capability to address complex challenges and facilitate intricate decision-making in several academic disciplines, resulting in significant advancements in innovation. The success of AI is associated with advanced technology, which has proven to impact our daily life activities significantly (Jena et al., 2021). AI has become more advanced due to the enhancement of many machine learning algorithms, particularly Deep Learning (DL), a novel form of Artificial Neural Network (ANN). The DL, a subfield of machine learning and artificial intelligence, draws inspiration from the human brain's ability to learn and extract meaningful information from data.

The role of DL has influenced many domains, such as industry and manufacturing, economics, healthcare and medical diagnosis, face recognition, robotics, internet applications and so on since it has excellent real-world applications (Górriz et al., 2020; Dang et al., 2021; Dong, Wang and Abbas, 2021; Nazir, Shakil and Khurshid, 2021). This phenomenon occurs because an enormous amount of data is produced every day using all kinds of devices with systems that store, accumulate and analyze in a structure that has been designed (Anagnostopoulos, Zeadally and Exposito, 2016). Through DL, AI has become more intelligent in extracting information from the data, with high-performance computational capabilities that can help design advanced decision-support systems. DL can efficiently handle massive volumes of high-dimensional data. Prediction classification is commonly applied in the medical and healthcare fields for early detection, diagnosis, prognosis, and treatment tasks. Early detection, diagnosis, and prognosis in diabetes research are significant areas of study. Moreover, Vu et al. (2020) have observed a substantial rise in publications focusing on the application of AI in diabetes, with several topics emerging as top areas of interest. They asserted that there are five prominent emerging research domains: (1) AI's applications in diagnosing diabetes; (2) Assessing the risk of diabetes and its complications; (3) The role of AI in developing new treatments and monitoring diabetes; (4) Utilizing telehealth and wearable technology for managing diabetes daily; (5) Analyzing the surgical outcomes of robotic procedures in patients with diabetes as a comorbidity. The following sub-section will discuss the motivation for risk prediction diabetes and complications with the prediction classification that is conducted in this study.

1.1.1 Motivation on Risk Diabetes Complication Prediction

Diabetes Mellitus (DM) or Diabetes is a group of metabolic disorders classified as chronic Non-Communicable Diseases (NCDs) that impact over 200 million individuals globally in the healthcare sector. Unhealthy food, obesity, lifestyle, genetics, lack of physical activity, and many other factors can cause this disease (Wing et al., 2001; Olokoba, Obateru and Olokoba, 2012). Diabetes is an intricate and highly significant illness, if not managed promptly and effectively, it can lead to severe complications, potentially resulting in death. There are several types of diabetes, including Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM), Gestational Diabetes Mellitus (GDM), Other Specific Diabetes and Diabetic Complications affecting organs like the heart, eyes, and kidneys (American Diabetes Association, 2018). However, T2DM commonly causes significant morbidity and mortality (Wang et al., 2013; Erkaymaz, Ozer and Perc, 2017). This is due to an increase in people with diabetes, which leads to complications mainly because of chronic hyperglycemia. There are two main objectives of current medication: saving lives and alleviating disease symptoms, preventing complications of long-term diabetes and eliminating various risk factors to enhance longevity (Kavakiotis et al., 2017). The high prevalence of diabetes is mainly caused by T2DM, which has increased by approximately 87% to 91% based on the estimation of the International Diabetes Federation (IDF) diabetes atlas 2015 (Kagawa et al., 2017). Every day, someone will be diagnosed with diabetes for 3 seconds and then a patient will die of complications due to diabetes for 7 seconds (Chen and Pan, 2018). T2DM sufferers are very dangerous if not handled properly and managed effectively. The estimated worldwide expenses for treating diabetes and its complications were 673 billion US dollars in 2015 and are projected to increase to 802 billion US dollars by 2040 (Ogurtsova et al., 2017).

However, the study indicates that diabetic patients commonly experience complications, with the most frequent being an increase in both microvascular and macrovascular issues. Besides, the total cost of management is up to 130% compared to those without complications. Diabetes with complications is a major health issue due to its significant impact on mortality, quality of life, and financial costs. Generally, the complications of diabetes mellitus are categorized into two types: microvascular and macrovascular, primarily resulting from hyperglycemia (Cade, 2008). Macrovascular complications include coronary artery disease, peripheral arterial disease, and stroke, while microvascular complications consist of diabetic nephropathy, neuropathy, and retinopathy (Fowler, 2008). Diabetes with complications can be categorized into acute complications such as diabetic ketoacidosis, hypoglycemia, diabetic coma, erectile dysfunction, respiratory infections, and periodontal disease. Chronic complications include heart failure, diabetic neuropathy, nephropathy, retinopathy, and diabetic foot, based on severity and time of onset