Deep Learning-based Decision Support Systems for Thyroid Cancer:

the Pathogenesis, Diagnosis, and Prognosis

Xinyu Zhang

Bachelor of Information Technology (Honours), Monash University

A thesis in the full fulfillment of the requirements presented for the degree of

Doctor of Philosophy

Department of Data Science and Artificial Intelligence

Faculty of Information Technology

Monash University

Supervised by Assoc/Prof. Vincent CS Lee, Dr. Jia Rong,

Assoc/Prof. James C Lee, Dr. Feng Liu

December, 2022

Copyright notice

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature:

Print Name: Xinyu Zhang

Date: 16 December 2022

Publications during enrolment

Journal articles:

1. Xinyu Zhang, Vincent CS Lee, Jia Rong, Feng Liu, and Haoyu Kong. "Multi-channel convolutional neural network architectures for thyroid cancer detection." Plos one 17, no. 1 (2022): *e*0262128. pp. 1-26. **Published**. *(SJR* 0.85*, Q*1 *Best Quartile, H-index:* 367*, Impact factor:* 3.752*)*

2. Xinyu Zhang, Vincent CS Lee, Jia Rong, James C. Lee, and Feng Liu. "Deep convolutional neural networks in thyroid disease detection: A multi-classification comparison by ultrasonography and computed tomography." Computer Methods and Programs in Biomedicine 220 (2022): 106823. pp. 1-10. **Published**. *(SJR* 1.33*, Q*1 *Best Quartile, H-index:* 115*, Impact factor:* 7.027*)*

3. Xinyu Zhang, Vincent CS Lee, Jia Rong, James C. Lee, Jiangning Song, and Feng Liu. "A multi-channel deep convolutional neural network for multi-classifying thyroid diseases." Computers in Biology and Medicine 148 (2022): 105961. pp. 1-11. **Published**. *(SJR* 1.31*, Q*1 *Best Quartile, H-index:* 102*, Impact factor:* 6.698*)*

4. Feng Liu and Xinyu Zhang. "Hypertension and Obesity: Risk Factors for Thyroid Disease." Frontiers in Endocrinology 1602 (2022). pp. 1-9. **Published**. *(SJR* 1.38*, Q*1 *Best Quartile, H-index:* 83*, Impact factor:* 6.055*)*

5. Xinyu Zhang, Vincent CS Lee, Jia Rong, James C. Lee, and Feng Liu. "Predicting thyroid cancer with machine learning: a comprehensive survey addresses current challenges and future opportunities." Nature Reviews Endocrinology. 43 pages. **In preparation for submission**. *(SJR* 8.56*, Q*1 *Best Quartile, H-index:* 165*, Impact factor:* 47.564*)*

6. Xinyu Zhang, Vincent CS Lee, Jia Rong, James C. Lee, and Feng Liu. "Integrating text mining with exceptional association mining: a comparative analysis on digital

health records." IEEE Journal of Biomedical and Health Informatics. 9 pages. **In preparation for submission**. *(SJR* 1.8*, Q*1 *Best Quartile, H-index:* 137*, Impact factor:* 5.772*)*

Conference articles:

1. Xinyu Zhang, Vincent CS Lee, Jia Rong, James C. Lee, and Qilin Zhang. "A Unified Model Selection Approach for Enhancements in Comprehensive Diagnostic Decision-making." **In preparation for submission**. *The* 26*th Pacific-Asia Conference on Knowledge Discovery and Data Mining. PAKDD,* 2023. 12 *pages. (Core A)*

2. Xinyu Zhang, Vincent CS Lee, Jia Rong, James C. Lee, and Feng Liu. "Faster Apriori: knowledge extraction from high dimensional digital records." **In preparation for submission**. *International Conference on Web Services (ICWS). IEEE,* 2024. 10 *pages. (Core A)*

Acknowledgement

I always remember the day when I got out of the ultrasonography room, when I couldn't even cry out loud. The pain engulfed me. I looked into my mom's eyes, and I saw sadness. I can recall my grievance that day, where I found myself isolated that no one else could feel the exact same pain and fear. That might be the motivation, well, at least one of the reasons I decided to go into depth about researching a new way to mitigate patients suffering from those anguished diagnostic procedures and my torturous memory.

Many say that differentiated thyroid cancer (DTC) can be seen as "good cancer" or even "lucky cancer" as patients with DTC usually have a promising prognosis. Well, this is what we usually hear and get comforted by, but this might be too optimistic. As a survivor, though I hate to call myself that, I'm grateful for the opportunities I had where I got to keep my voice and my good-looking. Not many people had the chance to make these decisions, they had to give up their voices, and some may even rely on machines to breathe. While I ended up here, at Monash, finishing my Bachelor's, Honours, and this, a Ph.D. degree. I'm lucky, and I'm so appreciative that I always get help from people around me to set me free from countless obstacles.

I would like to express my sincere appreciation to, my selfless supervisor, Associate Professor Vincent Cheng Siong Lee. Thank you very much for being so patient and caring with me. You always share your insights, wisdom, and knowledge with me and point the right direction for me when I'm confused or indecisive. You always guide me with warm advice, patience, and enthusiasm. You always take care of me like a family, you have inspired me on how to be an independent researcher, and also how to be a responsible, righteous, and considerate person. I would like to express my deepest appreciation for offering me numerous opportunities. And I will always miss our old times, having lunch together, having weekly meetings, and discussing histories and politics. Thank you so much, and I must say it is a great honour to be your $29th$ Ph.D. student.

I would also like to appreciate the guidance provided by my co-supervisors, Dr. Jackie Rong, Dr. James Lee, and Dr. Feng Liu. Thank you very much for supporting me through the hard times, offering me data sources and constructive comments. Thank you for sharing your expertise with me, it is your insightful and precious knowledge that complemented my research. Thank you for showing me what a strong work ethic is. I'm so grateful that you spent time discussing with me despite your busy schedule, and I'm fortunate to have you as my supervisors. A huge thanks to Dr. Feng Liu, I'm deeply thankful for saving my life. A special thanks to Dr. Jackie's family, you have immensely cared about me and provided me with much fabulous food, love, and joy.

I would like to thank my panel chairs, Associate Professor Arun Konagurthu and Associate Professor Guido Tack, and my panel members Dr. Lan Du and Dr. Lei Yang. Thank you very much for constantly reviewing my progress and providing insightful suggestions. Thank you for ensuring that I am on the right track to completion. It was nervous yet exciting to have the milestones during these years. Each time, you have offered me words of affirmation and made me a more confident person.

I would like to express my sincere gratitude to the participant patients, hospitals, and institutions which provided me with the research datasets. This research would not have been completed without their help.

I would like to extend my appreciation to our Graduate Research Communication and Academic Language Specialist, Ms Julie Holden. Thank you for organising various workshops to support developing our writing, communication, and research skills. Thank you for providing your expertise with patience in helping me improve my academic writing.

I would like to thank our Monash Graduate Research Office. Thank you for assisting me with the various administration processes on different occasions, such as enrolling, desk allocation, applying to study overseas, scheduling my milestones, and many more. Special thanks to Ms Rachael Unwin, thank you for your hard work in supporting my study life at Monash.

A special thanks to Associate Professor Jiangning Song and the Song Laboratory.

Thank you for being so sharing and always offering me opportunities to establish research connections and potential research directions. When I felt doubtful of myself, you always encouraged me, and I am so humbly grateful to have the chance to learn from you and incredible individuals from your lab.

I would like to thank Dr. Guanliang Chen for delivering warm suggestions and supporting me at all times. I want to thank you, my friends and colleagues, Minfeng Qi, Qilin Zhang, Haoyu Kong, Jionghao Lin, Van Nguyen, Zhikang Wang, and Ji Fang. Thank you for all you have done for me, and I will miss our days of work and joy.

Special thanks to my mentor during my Bachelor's industry experience project, Mr Harnam Thandi. You are always supportive and caring for me over the years. Thank you so much for encouraging me, which has truly been inspiring and motivating for me in going through this challenging journey. Australia does not make me feel that much lonely when you are there for me.

Thank you my supervisors during my Honours degree, Dr. Md Mahbubur Rahim and Dr. Sue Foster. Thank you for guiding me into the Ph.D. program and offering me the precious memory in doing research with you.

Thank you my friend, Yue Long. We knew each other since we were 11 years old, and you have always tolerated, cared about, and supported me in the past 15 years. Thanks to my friends Yanxi Long and Yashuang Deng, you have always encouraged and trusted me during the times when I felt disappointed in myself. A big thanks to Di Wang and Mingzhe Wu, it is grateful to have you during this journey.

Lastly, I want to express my sincere gratitude to my beloved family. You have gone through the entire process with me, and I know you always feel worried for me. I hope I made you proud. A special thanks to my mom who supports me both financially and mentally. You are the most lovely and brave human being I've ever known, and I hope I will be like you one day. I would like to thank my grandparents, dad, aunts, uncles, and cousins. Thank you for trusting in me. I would never have accomplished this degree without your unconditional support, encouragement, and trust.

For anyone who is about to read this thesis:

I was never a good writer that I couldn't make this thesis an intriguing piece of work, this is a professional document after all. And indeed, in the following chapters, I tried to make things with logic in a consistent and enjoyable way.

Hope you'd enjoy reading it.

Abstract

The thyroid, a butterfly-shaped endocrine gland locates at the base of the neck, is in charge of regulating the metabolic systems, including heart rates, blood pressure, and digestive functions. This crucial gland in the human body is now grabbing attention worldwide as thyroid cancer has been the fastest rising malignancy since 1982.

Despite the increasing number of instances, the pathogenesis of the disease remains unclear. The existing studies generally deploy qualitative or statistical techniques to investigate a single risk factor correlated with the development of thyroid cancer at a time. However, such an approach is inefficient and tends to ignore the connections among the diversified attributes, resulting in disagreement with the identification of thyroid disease risk factors.

The diagnosis of thyroid disease in the clinical domain is labour-costly and with varying degrees of uncertainty. The gold standard fine-needle aspiration cytology (FNAC) diagnosis heavily relies on the clinician's experience, leading to over 30% of the results being non-diagnostic. This effect directly aggravates patients' financial and physical suffering due to the increased missed diagnosis, unnecessary FNAC, or excisional biopsies rates. Along these lines, deep learning-based computer-aided diagnostic (CAD) systems incorporating medical images are arising as promising candidates for thyroid disease detection. However, diagnosing the disease subtype and co-existence phenomenon through CAD is neglected.

Additionally, the existing CAD systems generally use unitary datasets for training and cannot be adapted to different data sources, contributing to the limited clinical adoption. To address these challenges, this thesis contributes to 1) identifying risk factors correlated with thyroid disease, 2) enhancing diagnostic accuracy and efficiency, and 3) enriching clinical adoptions of deep learning approaches.

To determine thyroid disease pathogenesis, in Chapter 4, text mining and association rule mining techniques were adopted to extract common and exception rules simultane-

ously from raw patient medical reports. Extensive experiments were conducted to verify the identified risk factors based on different gender groups.

To enhance the diagnostic accuracy, Chapter 5 incorporates deep convolutional neural networks (CNN) with pre-operation medical images to achieve accurate diagnosis from binary and multi-class classification tasks. In Chapter 6, three multi-channel CNN architectures were developed to achieve a comprehensive diagnosis for the entire thyroid gland, reaching patient-specific diagnoses meanwhile considering the disease subtype co-existence phenomenon. The experimental results demonstrate unprecedented diagnostic performance and generalisation to different gender groups.

To elevate the potential applications of CAD systems in the clinical domain, a unified model selection approach was proposed in Chapter 7, which can be adaptive to different data sources with distinct patient profiles. The approach consists of a self-directed individual network selection mechanism, a dynamic weighting scheme, and a weighted ensemble averaging model, tailored to generate robust and reliable diagnostic decisions. The experimental results demonstrate that the approach can reach satisfying performance under different scenarios.

In summary, the investigations conducted in this thesis revealed promising performance in the experiments and ablation studies with a comprehensive evaluation metric, including area under the curve (AUC), accuracy, precision, recall, specificity, negative predictive value (NPV), false-positive rates (FPR), and F1 score. With the help of deep learning techniques, this thesis is dedicated to understanding thyroid disease and enhancing diagnostic accuracy and efficiency for a promising prognosis.

Keywords: *Deep learning, Machine learning, Thyroid Cancer, Association rule mining, Multi-channel convolutional neural network*

Contents

List of Figures

List of Tables

Abbreviations

ACWE	**A**ctive **C**ontours **W**ith **E**dges
AMF	**A**daptive **M**edian **F**iltering
ANN	**A**rtifical **N**eural **N**etwork
ARM	**A**ssociation **R**ule **M**ining
ATC	**A**naplastic **T**hyroid **C**arcinoma
AUC	**A**rea **U**nder the **C**urve
BMI	**B**ody **M**ass **I**ndex
CCE	**C**ategorical **C**ross **E**ntropy
CNN	**C**onvolutional **N**eural **N**etwork
CPIR	**C**onditional **P**robability **I**ncrement **R**atio
CT	**C**omputed **T**omography
CV	**C**ross **V**alidation
DDTI	**D**igital **D**atabase of **T**hyroid Ultrasound **I**mages
DIDC	**D**ouble **I**nputs **D**ual **C**hannel
DM	**D**ata **M**ining
DSC	**D**ice **S**imilarity **C**oefficient
DT	**D**ecision **T**ree
DTC	**D**ifferentiated **T**hyroid **C**ancer
EACCD	**E**nsemble **A**pproach for **C**lustering **C**ancer **D**ata
EDA	**E**xploratory **D**ata **A**nalysis
FCM	**F**uzzy **C**ognitive **M**ap
FCN	**F**ully **C**onvolutional **N**etwork
FNAC	**F**ine **N**eedle **A**spiration **C**ytology
FPR	**F**alse **P**ositive **R**ate
FPT	**F**requent **P**attern **T**ree
FS	**F**eature **S**election

FT3	**F**ree **T**riiodothyronine
FT4	**F**ree **T**hyroxine
FTC	**F**ollicular **T**hyroid **C**arcinoma
GMM	**G**aussian **M**ixture **M**odels
KNN	**K** **N**earest **N**eighbour
KW	**K**ruskal **W**allis
LR	**L**ogistic **R**egression
LSTM	**L**ong **S**hort **T**erm **M**emory
MI	**M**utual **I**nformation
MIL	**M**ultiple **I**nstance **L**earning
MLP	**M**ulti **L**ayer **P**erceptron
MRI	**M**agnetic **R**easoning **I**mage
MTC	**M**edullary **T**hyroid **C**arcinoma
NB	**N**aive **B**ayes
NPV	**N**egative **P**redictive **V**alue
PPV	**P**ositive **P**redictive **V**alue
PTC	**P**apillary **T**hyroid **C**arcinoma
RF	**R**andom **F**orest
RNN	**R**ecurrent **N**eural **N**etwork
ROI	**R**egion **O**f **I**nterest
RQ	**R**esearch **Q**uestion
SIDC	**S**ingle **I**nput **D**ual **C**hannel
SVM	**S**upport **V**ector **M**achine
TIRADS	**T**hyroid **I**maging **R**eporting and **D**ata **S**ystem
TM	**T**ext **M**ining
TSH	**T**hyroid **S**timulating **H**ormone

Notations

Symbol	Descriptions	Constraints
$\mathcal{D}$	Dataset	
$\bar{\mathcal{D}}$	Updated dataset	
$\mathcal{R}$	Real numbers	
X	Encoded instance matrix	$X = \{X_1, X_i, X_n\}$
y	Encoded instance label	$y = \{0, 1\}$ or $y = \{0, 1, \ldots, 5\}$
$\mathcal{N}$	Number of instances	
n_w	Feature map width	
n_h	Feature map height	
n_c	Feature map channel number	
m	Individual attribute from instance X_i	
w	Weight	$0 \leq w \leq 1$
$\neg$	Exception rule	
θ	Threshold value	$0 < \theta < 1$
s	Stride number	
f	Convolutional kernel size	
b	Bias	
K	Number of folds for CV	$K = 5$ or $K = 10$
C	Number of classes	$C = 2, 6, 16$
$\mathcal{T}$	Trained CNN	
$\mathcal{L}$	Performance metrics	
$\mathcal{P}(\mathcal{C}\|\mathcal{A})$	Conditional probability of $\mathcal{A}$ given $\mathcal{C}$	$0 \leq \mathcal{P}(\mathcal{C}\|\mathcal{A}) \leq 1$

Chapter 1

Introduction

The thyroid is considered the largest endocrine gland in an adult, shaped like a butterfly and located at the lower neck [104]. It controls the daily metabolism of cells by producing hormones and regulating the balance of calcium in the human body, including heart rate, body temperature, and bloodstream velocity [213]. With an adequately functioning thyroid gland, one can maintain the right amount of hormones to keep metabolism activities at a favourable rate [267]. However, improper functioning of the gland will lead to thyroid-related diseases. For this reason, this crucial gland has gradually attracted considerable attention in the medical and technology domains since it gives rise to the most common endocrine tumours [25].

Thyroid diseases are highly pervasive and can be broadly classified into functional and neoplastic kinds, while they can co-exist [169]. Functional thyroid diseases mainly include hypothyroidism, hyperthyroidism, Graves disease, Hashimoto's disease, and thyroiditis [320]. Neoplastic diseases can develop into tumours (i.e., benign and malignant) [320]. In addition, thyroid tumours are the most commonly seen nodular lesions among adults [89]. More specifically, over 50% of adults have thyroid nodules [3]. Most of these nodules are benign, such as cystic and adenoma [320]. However, 5% to 15% of the thyroid nodules are malignant, which will lead to the formation of the four types of thyroid cancer: papillary thyroid carcinoma (PTC), follicular thyroid carcinoma (FTC), medullary thy-

roid carcinoma (MTC), and anaplastic thyroid carcinoma (ATC) [224]. Thyroid cancer is typically painless and usually undetectable by patients themselves [224].

Since the $20th$ century, thyroid cancer instances have progressed, rising at the fastest rate among all the malignancies [303]. According to the latest Cancer Fact & Figures statistics, there will be an estimation of $43,800$ new thyroid cancer instances diagnosed in the United States of America by the end of 2022, with a death rate of 4.62% [283]. Thyroid cancer instances are rapidly increasing, with the highest rates found in the Federated States of Micronesia, French Polynesia, North America, and East Asia [247]. In particular, the highest incidence-to-mortality rates have been reported in South Korea, Cyprus, and Canada [247]. In China, thyroid cancer was ranked as the $4th$ most commonly diagnosed malignancy in females [81]. Among the established instances, female patients are three times more likely to develop the disease than male patients [283].

With the implication of the increased morbidity and mortality rates brought by thyroid cancer, challenges in understanding the cause and gender disparity, enhancing the diagnostic performance, and targeting promising prognosis, are all desired to be addressed. Accordingly, this research sought to provide a structured organisation of the existing studies, determine the underlying cause of thyroid cancer, and strive for a more efficient procedure to reach a precise diagnosis using machine learning (i.e., data mining, deep learning) techniques. With these techniques, clinicians can gain a "second opinion" relying on decisions made by computers to focus more on patient care, enabling promising prognoses for patients with thyroid cancer.

1.1 Background and Motivation

Differentiated thyroid cancer (DTC) is regarded as "good cancer" or even "lucky cancer" by society, where patients commonly confront the perception that their prognosis is relatively promising [255]. DTC consists of PTC and FTC kinds [39], and usually has relatively more favourable survival rates, being more than 95% for the five-year survival

[155]. Patients with DTC normally expect to have an average lifespan after treatments (i.e., thyroidectomy and radio-therapy). Nevertheless, rarer types, such as ATC or MTC, have much lower survival rates, being less than 10% for the five-year survival [204]. Clinicians usually emphasise the optimistic outcomes for comforting patients. Unfortunately, this inadvertently weakens the impacts on patients' lives brought by thyroid cancer [255].

There exist long-established protocols for identifying, diagnosing, and treating thyroid cancer in the clinical setting. Although it is considered a pervasive disease, there are still many challenges in understanding and diagnosing thyroid cancer in the clinical domain. Accordingly, the following context illustrates the existing clinical protocols for the pathogenesis, diagnosis, and prognosis of thyroid cancer, interprets current challenges clinicians face, and then emphasises the motivations of this research work.

1.1.1 Thyroid Cancer Pathogenesis

For decades, academia and medical fields have always been struggling to identify the leading cause, gender disparity causes, and the recurrence causes of thyroid cancer, as the pathogenesis is affected by multiple diversified attributes. Some scholars believe that family history will increase the risk of establishing thyroid cancer [326, 18], while others argue that poor lifestyle [190], mental health issues (i.e., depression) [154], periodical health conditions [10], and environmental factors (e.g., air pollution) [136] might also play crucial roles. In the past few decades, the scientific community has been dedicated to identifying risk factors associated with thyroid cancer, which allows to potentially reduce morbidity and mortality rates.

The intrinsic examination process is that clinicians independently select one risk factor and evaluate its association with thyroid cancer development. Following this process, several potential factors (including comorbidity) were identified correlated with thyroid cancer, including radiation exposure [134, 31, 292, 26], iodine intake level [360, 344, 137], dietary nitrate intake level [314, 124], vitamin D deficiency [138], diabetes [299, 334, 40], obesity [325, 184, 266, 329], gene heredity and mutation [107], smoking status [56], family

history [62], and hormonal factors [202, 194, 116].

However, plenty of the identified factors are still controversial as they were generally analysed with different groups of patients, such as different scales of the study groups or diversified demographic features. Therefore, no consensus can be established on the investigated factors. More importantly, the evaluations of either one of those factors take long-established retrospective investigations, let alone the interwoven among them. Ignoring the correlations among diversified risk factors can also bring substantial errors in understanding the cause of thyroid cancer. Therefore, there has been considerable debate in the literature around the identification of these risk factors and there is still limited research to fully understand their interpretability and reliability.

In summary, the associations among the multiple risk factors cannot be evaluated by existing studies, thus amplifying the misinterpretations in understanding thyroid cancer causes. Confirming the associations between the individual factors and thyroid cancer development is quite challenging and inefficient, rendering thyroid cancer risk factors identification tasks barely achievable. These shortcomings create limitations for revealing the pathogenesis of thyroid cancer. Therefore, this research needs the means to incorporate an automated algorithm to efficiently mine essential knowledge from high-dimensional medical records to reveal the pathogenesis of thyroid cancer.

1.1.2 Thyroid Cancer Diagnosis

With respect to the dramatic growth rate of thyroid cancer instances, it is mainly owing to the advanced sensitive detection procedures [283]. The clinical diagnostic procedure for thyroid cancer is standardised. By following the clinical guidance, the thyroid function examination is always in priority, which measures hormones produced by the gland, including thyroid-stimulating hormone (TSH), triiodothyronine ($T3$), thyroxine ($T4$), free triiodothyronine ($FT3$), and free thyroxine ($FT4$) [144]. TSH is considered the most sensitive parameter for detecting thyroid dysfunction, while $FT3$ and $FT4$ are supportive parameters for monitoring therapies. Functional thyroid disease can usually

be diagnosed through thyroid function examinations [65].

A patient who has undergone health examinations identifying abnormal thyroid nodules via medical images would need to perform fine-needle aspiration cytology (FNAC) to further evaluate the problematic nodules. Ultrasonography (ultrasound) is the most well-accepted medical image modality for thyroid imaging on suggestive of malignancy as it is safe, non-invasive, and cost-efficient [289]. Ultrasound provides an intermediate risk for patients following the *Thyroid Imaging Reporting and Data System* (TIRADS) score based on the features appearing on the ultrasound images, such as irregular margins, hypo-echogenicity, taller-than-wide shape, and microcalcifications [85]. TIRADS scores from 1 to 5 where 1 was defined as "*normal*", 2 as "*benign*", 3 as "*no suspicious features*", and 4*a* as "*one suspicious feature*", 4*b* as "*two suspicious features*", 4*c* as "*three to four suspicious features*", and 5 as "*five or more suspicious features*", respectively [19]. Nevertheless, TIRADS scores cannot provide precise decisions for thyroid cancer diagnosis. Based on the analysis conducted by American Thyroid Association [19] where 951 patients' TIRADS scores were compared to the ground-truth histopathological results. The authors found that thyroid cancer was appearing at a 0.9% with TIRADS ranking 2, 2.9% cancer rate was found in TIRADS 3, TIRADS 4*a* had a cancer rate of 12.3%, 34.4% cancer rate was found in TIRADS 4*b*, 66.6% in TIRADS 4*c*, and 86% cancer rate was in TIRADS 5. Deriving from this finding, solely relying on ultrasound images to make a diagnosis through TIRADS scores is considered insufficient. In this scenario, patients having TIRADS scored from 2 to 5 usually need to perform further FNAC to determine whether the nodule has cancerous cells [286].

FNAC is regarded as the gold standard in assessing the malignancy of thyroid nodules [24]. It gets biopsies from suspicious nodule cells using the apparatus following the ultrasound guidance, then has the biopsies assessed by pathologists to make diagnostic decisions regarding nodule malignancy risk stratification. Figure 1.1 demonstrates the FNAC apparatus and Figure 1.2 displays the implementation procedures. FNAC examinations heavily rely on the clinician's professionalism and expertise. Due to the lack

of experience of pathologists, over 30% of FNAC results turn out to be non-diagnostic, inaccurate, or indeterminate [234]. Patients with indeterminate FNAC results usually need another FNAC or excisional biopsy to have the nodule removed just for diagnostic decision-making. Sometimes patients might undergo unnecessary surgeries, and then determining the nodule shows no evidence of malignancy [72].

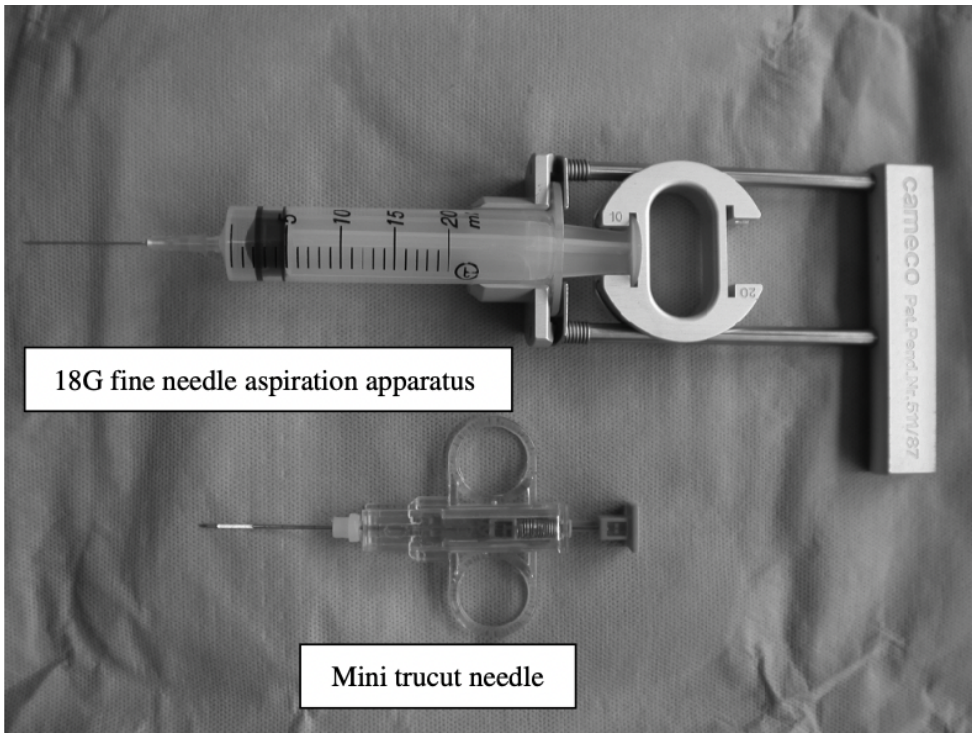


Figure 1.1: FNAC apparatus (sample image from [224]).

Under this rigorous circumstance, the clinical diagnostic procedure for thyroid cancer is tedious and anguished, arousing emotional, as well as financial burdens for patients. Due to the human false-positive and false-negative rates, the diagnostic decision is relatively subjective, and the clinical process is inefficient. These shortcomings create limitations for thyroid cancer diagnosis in the clinical setting. Therefore, clinicians seek to improve diagnostic accuracy and efficiency by integrating human knowledge with computer-driven techniques. With the emergence of machine learning, particularly the deep learning concept, the diagnostic procedures form into advanced ways of making decisions, named "computer-aided diagnosis (CAD)" techniques.

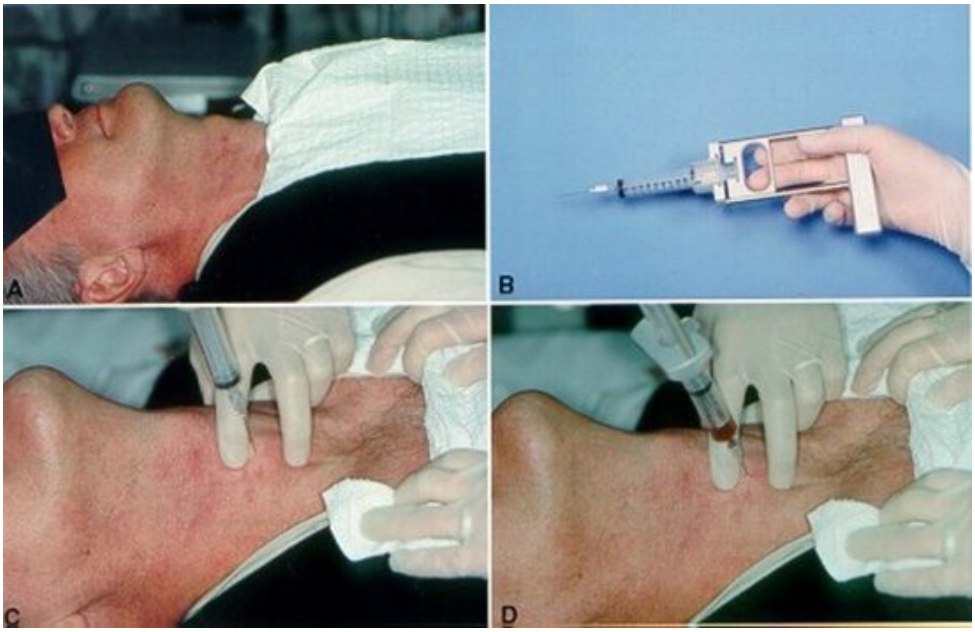


Figure 1.2: FNAC procedures (sample image from [66]).

1.1.3 Thyroid Cancer Prognosis

Differentiated thyroid cancer follows a standardised treatment protocol which usually requires surgery, such as partial thyroidectomy, total thyroidectomy, or total thyroidectomy with lymph node dissection [268]. Further radiation therapy is expected to mitigate recurrence risks by performing radioactive iodine (i.e., Iodine-131 treatment) or observation depending on any signs of metastasis [41]. Patients with thyroidectomy are expected to take levothyroxine one hour before breakfast for thyroid hormone replacements in their lifetime [307]. Levothyroxine is a manufactured form of the thyroid hormone thyroxine, and the common thyroxine medications are Eutroxsig, Oroxine, and Aspen Pharma [307].

Jegerlehner et al. [121] suggested that a substantial and growing part of the detected thyroid cancers were over-diagnosed and over-treated. The post-therapy hypothyroid disease frequently occurs after surgery or Iodine-131 treatments [12]. The utilisation of radioiodine therapy would somehow bring adverse effects for patients [269]. Lee et al. [157] also confirmed that Iodine-131 treatment would cause salivary gland dysfunction if the dosage of iodine was not carefully determined. Since each patient would undergo different disease stages, customised treatment plans are deemed crucial to avoid over-treatments by

offering them customised treatment plans based on age, weight, disease stage, medication intake level, medical history, and comorbidity, to name a few. Customised treatments can also help patients to understand their disease status more explicitly, and clinicians can gain supportive opinions on targeted treatments besides guidelines. The promising prognosis of thyroid cancer can be obtained only if the appropriate treatments are offered. To achieve this goal, a customised treatment decision support system is regarded as necessary.

Besides providing customised treatment plans, predicting the recurrence and survival rates can also potentially improve the prognosis for thyroid cancer patients. In reality, patients diagnosed with rare thyroid cancer tend to have a considerably shorter lifespan than patients diagnosed with DTC. The prognosis of patients with rare thyroid cancer types is relatively unsatisfying due to the higher recurrence rates and death rates. Among the four types of thyroid cancer, PTC has the highest survival rate. On the contrary, rarer kinds have much lower survival rates, which is less than 10% [204]. Mazzaferri and Jhiang [199] once indicated that 20% of thyroid cancer patients would experience recurrence, in which they might need re-operations or repetitive radiotherapy treatments. Unlike the pathogenesis of thyroid cancer, the risk factors correlated with thyroid cancer recurrences are determinate, including gender, elder age, primary disease extent, metastases to other organs, tumour size, extra-thyroidal invasion, location of nodules, and cervical lymph nodules [281, 33, 244, 93, 126, 211]. Wang et al. [311] proposed that PTC patients who had a total thyroidectomy and neck dissections tend to have a 3.8% recurrence rate. In addition, the various surgical treatments also differentiate recurrence rates of thyroid cancer. Based on the comparative study by Zetoune et al. [339], the recurrence rate for patients with thyroidectomy and neck dissection is 2.02%, whereas the recurrence rate for patients with thyroidectomy only is 3.92%. It is evident that thyroidectomy and iodine treatment decrease the recurrence rate of thyroid cancer [199, 244, 302].

Collectively, thyroid cancer prognosis is closely related to well-established treatment protocols. Therefore, a thyroid-specific treatment recommendation system, which offers

customised treatment plans for individual patients based on their health conditions and preferences, is highly desired. Moreover, predicting survival, death, and recurrence rates will potentially help achieve a promising prognosis for thyroid cancer patients.

1.2 Research Objectives and Questions

Even though clinicians strive to comfort patients by advocating that thyroid cancer has high remission rates, they should precept that patients are mentally vulnerable. Therefore, clinicians should seek practical strategies for achieving effective diagnostic procedures and promising prognoses to assist patients.

Under this direction, machine learning-based techniques have demonstrated promising performance in addressing challenges around understanding, diagnosing, and treating various types of diseases [54, 28, 215, 207]. These CAD techniques harnessed the power of computers to automatically learn from a large scale of experiences and provide precise decisions around the disease. The application of such advanced CAD techniques should be adaptive to diversified medical cultures to encourage them in the clinical setting for implementation to mitigate human false-positive and false-negative risks while offering clinicians second opinions. This potentially helps to detect disease more accurately and alleviates the possibility of over-treatment. Accordingly, this thesis applies machine learning approaches to reveal the unknown of thyroid disease, majorly thyroid cancer, leveraging comprehensive investigations around its pathogenesis, diagnosis, and prognosis. Three objectives for addressing three research questions (RQ) are as follows:

- **Objective** 1: To gain insights for a more comprehensive and deterministic understanding of the pathogenesis of thyroid disease through the utilisation of data mining techniques. Data mining techniques are tailored to extract valuable knowledge from complex attributes. Investigating the associations between the risk factors and thyroid disease development contributes to understanding the epidemiology. Therefore, a mixed method was deployed in this thesis by integrating qualitative and quantita-

tive analysis to determine the pathogenesis and gender disparity of thyroid disease (i.e., disorder and cancer).

  - ***Research Question*** 1: *What pathogenesis of thyroid disease can be ascertained through data mining approaches?*

- **Objective** 2: To diagnose thyroid disease in an advanced manner with enhanced efficiency and reduced human false-positive and false-negative rates through deep learning techniques, highlighting the importance of multi-classifying thyroid disease types and diagnosing disease co-existence scenarios. Therefore, this thesis adopts deep convolutional neural networks (CNN) to provide clinicians assistance regarding precise thyroid disease diagnostic decision-making, which mitigates patients' physical and financial pressure caused by clinical diagnosis.

  - ***Research Question*** 2: *How can the diagnosis of thyroid disease be improved via deep learning techniques?*

- **Objective** 3: To build a unified model adaptable to different patient profiles with a dynamic weighting scheme based on pre-determined performance criteria, tailored for precise thyroid cancer detection tasks. Incorporating data from diversified sources helps to build a unified model, which utilises limited information in generating comprehensive and reliable diagnostic decisions. Therefore, this thesis sought to design a general model that selects the best-performing individual networks under diversified sources and assembles them into a robust system that can be generalised to different institutions. This allows to elevate CAD applications in the clinical domain and helps establish a well-designed treatment decision support system for improving the prognosis of this disease.

  - ***Research Question*** 3: *How can a deep learning-based model be adaptive to different patient profiles for thyroid cancer diagnosis?*

1.3 Research Contributions

This thesis is dedicated to making contributions both theoretically and practically, presented from the following perspectives.

1.3.1 Theoretical Contributions

- **Contributing to add incremental values to the community.** This research proposes a comprehensive literature review framework, which can be adapted to different research scenarios for knowledge extraction. Following the established framework, a structured organisation of extensive existing research applications was described in this work, exhibiting current research challenges and corresponding future research opportunities (in Chapter 2).

- **Enhancing potential applications of CAD-based techniques in the clinical domain.** Machine learning-based approaches yield promising performance and efficiency for disease detection and treatments, yet their implementation is sobering. This research then shifts the focus to help implementation and the practical issues of actionality of deep learning-based techniques in the clinical domain [273, 301], contributing to the potentially enhanced clinical adoptions of the CAD-driven techniques (in Chapters 3 - 7).

1.3.2 Practical Contributions

- **Determining thyroid disease risk factors helps to mitigate morbidity and mortality rates.** With data mining techniques, the pathogenesis of thyroid cancer can be revealed, and the identified risk factors can be stratified based on their associative rankings. Accordingly, clinicians can work towards mitigation strategies to counter those dominating factors, minimising the side effects and eventually benefiting society from the public health domain (in Chapters 3 and 4).

- **Thyroid disease diagnosis can be much more accurate and effective.** This thesis incorporates deep learning techniques with two pre-operative medical imaging modalities for thyroid disease detection, emphasising the importance of multi-classifying the disease subtypes and co-existence situations. Through the proposed deep learning models, automatic diagnosis of thyroid disease can be achieved in an advanced, accurate, and efficient manner. Moreover, this thesis sought to more closely emulate the human-level diagnostic process to assist clinicians with offering preliminary decisions. This research is expected to reveal whether deep learning approaches have the potential to replace human diagnosis. The proposed CAD systems can also be further adapted to different diseases, enhancing the clinical applications (in Chapters 3 and 5).

- **Enhancing the use of patient-specific designs in CAD implementations.** This thesis proposes decision support systems, which were built following patient-specific design for the diagnosis and potentially for the treatments. Therefore, the models can make decisions for individual patients at a time (in Chapters 3 and 6).

- **Generating robust, comprehensive, and reliable decisions with limited information.** A unified model selection approach is proposed, which selects individual networks pre-trained with cross-institutional data sources under different medical cultures. The approach dynamically assigns weights based on the individual networks' pre-determined performance criteria to generate the weighted ensemble averaging model, enhancing its generalisation to diverse data populations tailored for precise thyroid cancer diagnosis (in Chapters 3 and 7).

- **Achieving promising prognosis for patients with thyroid disease.** Patient-specific design for diagnosis and treatments leads to enhanced prognosis by offering a customised treatment recommendation system for individual patients. It allows clinicians and patients to understand his/her health status more explicitly and offers customised treatment protocols based on the patient's preferences. In this regard,

patients can pay close attention to the health condition changes at any moment to obtain an improved prognosis. In the meantime, the survival rates can be prolonged, and death/recurrence rates can be minimised (in Chapters 3 and 8).

1.4 Thesis Structure

The systematic structure of this thesis is presented in Figure 1.3, and the main contents of each chapters are summarised as follows:

- Chapter 1 interprets the background, motivations, and objectives behind this research project around the pathogenesis, diagnosis, and prognosis of thyroid cancer.
- Chapter 2 briefly reviews all the related works and identifies literature gaps from the proposed framework.
- Chapter 3 elaborates the overall research paradigm, methodological design, and data acquisition for the defined three research questions.
- Chapter 4 identifies thyroid disease-associated risk factors through data mining techniques. The adopted association rule mining and feature selection algorithms further verify the identified factors.
- Chapter 5 presents the increased diagnostic efficiency and accuracy of thyroid disease through deep convolutional neural networks (CNN) incorporating medical images from binary and multi-classification tasks.
- Chapter 6 proposes various multi-channel CNN architectures for detecting thyroid disease subtype co-existence situations and generalising to different gender groups.
- Chapter 7 describes the proposed unified model, which can be adapted to different data sources with distinct patient profiles under different clinical scenarios, enriching CAD applications.

- Chapter 8 concludes the thesis and summarises the salient findings. Future research directions are also discussed.

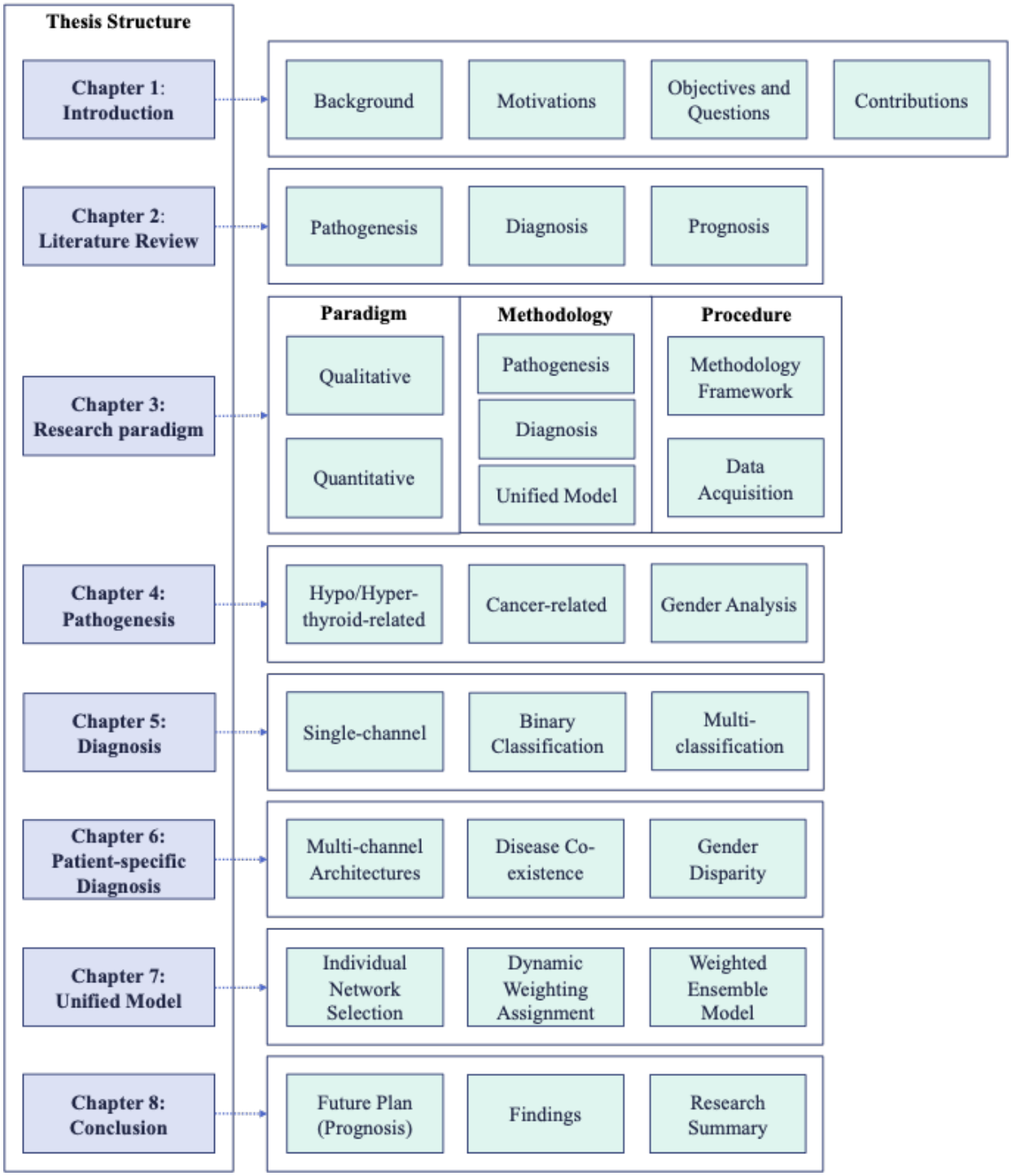


Figure 1.3: Thesis systematic structure.

Chapter 2

Literature Review

In order to conduct a comprehensive systematic literature analysis, a detailed step-by-step literature review framework was proposed in this thesis, and the framework is presented in Figure 2.1.

The systematic literature review framework contains the following three stages: research definition, research methodology, and research analysis. "*Research definition*" includes identifying the research field, defining research objectives, and outlining the research scope. "*Research methodology*" oughts to identify related literature studies based on the pre-defined searching strategies. "*Research analysis*" analyses the identified studies, presenting the key findings and interpreting the existing literature gaps.

2.1 Research Definition

The first stage of the literature review framework is the research definition. The research focus has been identified in this phase. Meanwhile, the research motivations and objectives were highlighted, and the research scope was defined.

This research seeks to identify machine learning-related studies to identify thyroid disease risk factors and help reach automatic diagnoses and promising prognoses. Therefore, this thesis aims to analyse literature to help comprehend, diagnose, and treat thyroid cancer through machine learning, majorly deep learning techniques.

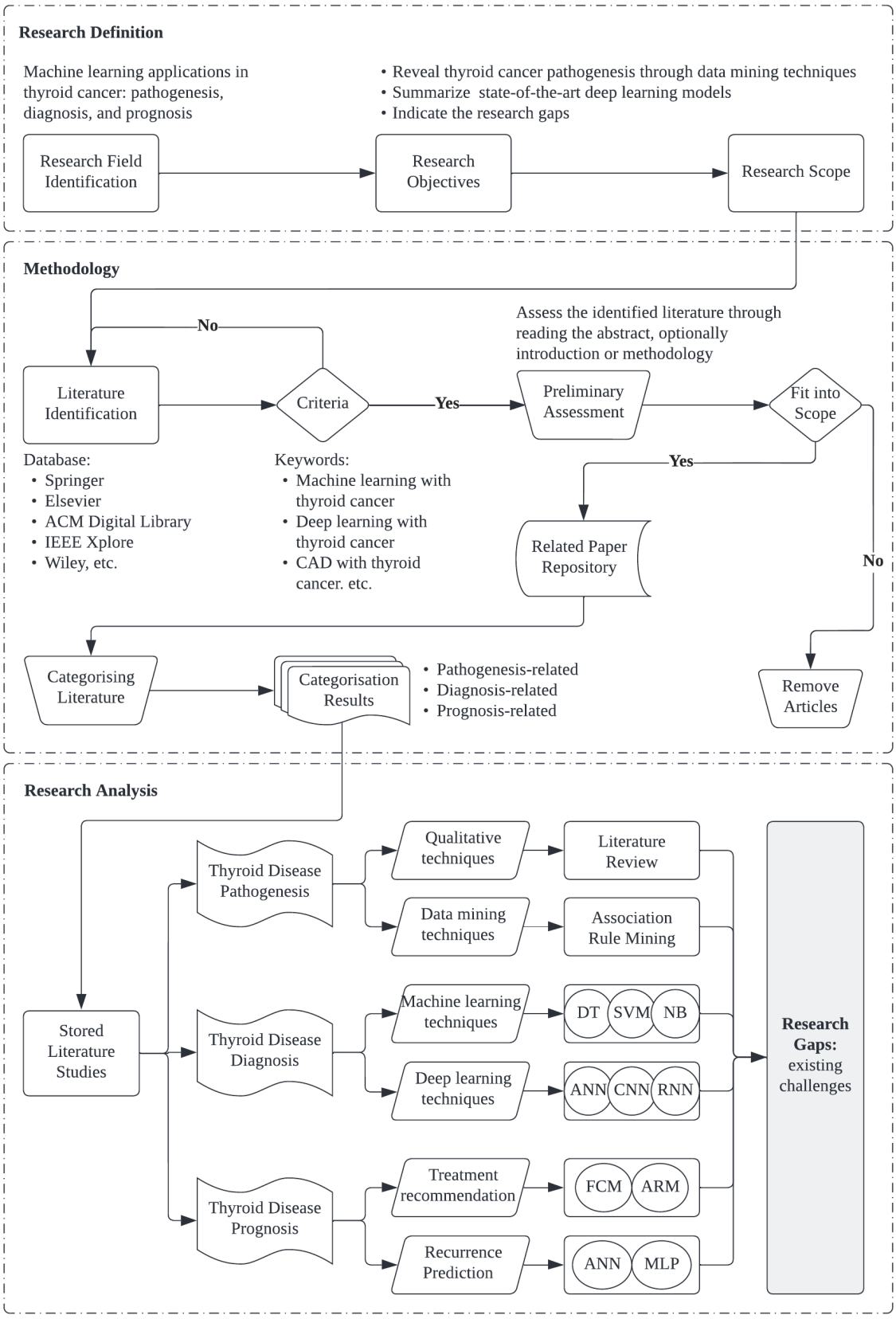


Figure 2.1: Systematic literature review framework.

Lastly, the scope of this thesis is to include as many related literature works as possible for analysis to establish research gaps that immensely fill clinical gaps. Machine learning techniques have been applied relatively often in the medical field, whereas the application of those techniques in the thyroid domain is still limited. In order to have a comprehensive analysis, this thesis intends to include as many high-level ranking literature studies as possible. Moreover, the specific literature identification process will be explained in the next section.

2.2 Research Methodology

After defining the research scope, a set of literature identification processes was conducted. This research has included a list of high-level ranking conference and journal databases during the literature searching phase, such as Elsevier, Springer, IEEE Xplore, and ACM Digital Library, to name a few. A set of criteria were used as our search protocols, and Table 2.1 lists some keywords applied during the searching phase.

Table 2.1: Literature identification searching phase.

Research Focus	Searching Keyword / Phases
Pathogenesis	Risk factors of thyroid disease
	Thyroid disease pathogenesis with machine learning
	Thyroid disease pathogenesis with data mining
	Thyroid disease risk factors with association rule mining
Diagnosis	Machine learning with thyroid disease
	Deep learning with thyroid disease
	CAD for thyroid disease detection
Prognosis	Thyroid treatment recommendation system
	Machine learning with thyroid disease treatment
	Prediction of thyroid disease survival or recurrence

During the searching phase, no timeline restriction was followed to explore the developments of advanced machine learning approaches. In this regard, the literature studies were selected based on the flowchart (Figure 2.2) after applying the above-identified keywords in the academic database.

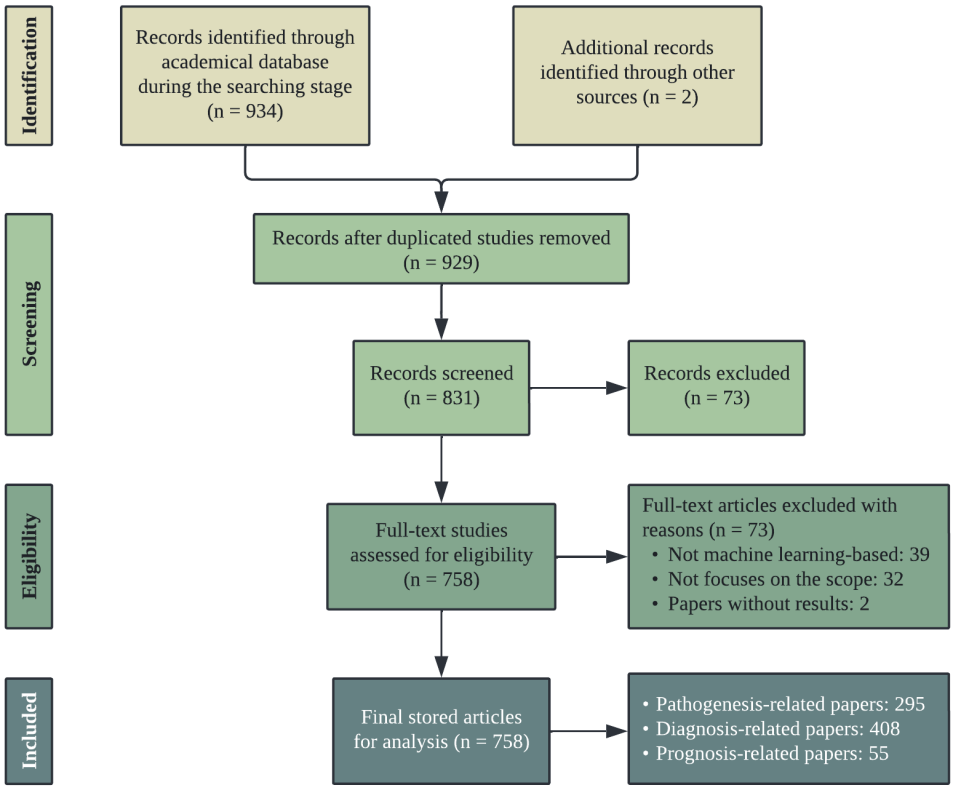


Figure 2.2: Flowchart of the searching strategy and literature selection.

The abstract of each paper has been analysed to evaluate its relevance to the defined scope. The methodology section of the paper was optionally analysed if the abstract was not explicit enough regarding the utilisation of machine learning techniques. The papers that fit into the research scope were stored in the literature repository for further analysis, otherwise were removed. Lastly, all the identified papers were categorised into three groups, namely the pathogenesis, diagnosis, and prognosis of thyroid disease. As a result, 295 papers were categorised into the pathogenesis group, 408 papers were categorised into the diagnosis group, and 55 papers were categorised into the prognosis group.

2.3 Research Analysis

A total of 758 papers were utilised for the analysis phase of this thesis. Each of the three categories was analysed based on their performance, feasibility, and generalisation of the proposed machine learning approaches.

The thyroid disease pathogenesis-related studies were divided into qualitative and quantitative investigation groups. The qualitative studies investigated risk factors correlated with thyroid disease, while the quantitative studies confirmed some risk factors through statistical and data mining techniques. For thyroid disease diagnosis-related studies, deep learning procedures were discussed independently. As far as thyroid cancer prognosis-related studies are concerned, the design of treatment recommendation systems and the prediction of survival, death, and recurrence rates were systematically explained. Collectively, a summary of the literature analysis will be presented, existing challenges inferred from literature analysis will be outlined, and the corresponding research questions will be re-addressed.

2.3.1 Pathogenesis and Risk Factors

Thyroid cancer pathogenesis has always been on hit in the clinical and academic domains. Qualitative and quantitative analyses were usually conducted to reveal the origins of such a mystery. Qualitative literature was conducted based on survey analysis, and quantitative studies were majorly built based on statistical learning, such as case-control, retrospective, or prospective cohort studies. By investigating the identified 295 pathogenesis-related studies, seven risk factors were extracted and examined.

Radiation

The non-debating factor causing thyroid cancer is radiation exposure, such as medical or ionizing. The pediatric thyroid gland is regarded as the most sensitive organ to radiation [128]. Different kinds of radiation expose potential risks for establishing thyroid cancer, especially during childhood. In particular, neck exposure to ionizing radiation highly increases the risk of developing thyroid cancer at an early age, sometimes can be affected by medical scanning and external radiation [315, 271, 6, 245, 319]. Accordingly, radiation exposure at an early age is significantly relevant to thyroid cancer development.

Gene Heredity and Mutation

Gene heredity is considered another consensus-established risk factor associated with thyroid cancer. Gene heredity is to inherit chromosome pairs from families [59]. Based on the interview conducted by Ito et al. [118], 5% of PTC instances are inherited from family. Furthermore, gene mutation has clearly been announced to be highly relevant to thyroid cancer pathogenesis [235], which is mainly responsible for MTC development [294, 223, 118]. Many studies have been conducted to identify specific gene mutation types causing thyroid cancer [82, 107].

Iodine Intake

The most well-known substance which has an impact on thyroid glands is iodine. Typically, high-level iodine foods in our daily life are dairy products like milk, eggs, seafood, marine products (e.g., fish, seaweed, and shrimp), and iodised salt [69]. Over the past few decades, the scientific community has been dedicated to determining the association between the iodine intake level with the thyroid gland. Michikawa et al. [205] found that excessive seaweed consumption could lead to thyroid cancer, especially for postmenopausal women. Zimmermann and Galetti [360] reported that iodine deficiency is a risk factor for thyroid cancer development, particularly for the follicular and possibly for the anaplastic kinds, through meta-analysis. The thyroid gland is sensitive to low intake of iodine has been confirmed by several other studies [118, 315]. Overall, iodine is undoubtedly a risk factor associated with thyroid cancer regardless of its excessive or limited intake, while today, the mechanism of their linkage is still unknown.

Dietary Nitrate Intake

With the rapid development of infrastructures and industries, more and more environmental pollutants generated and influenced our daily lives. Among all the pollutants, nitrate is a contaminant of drinking water, and dietary nitrate is a kind of component of daily diets that occurs at a high-level within green leafy and root vegetables [314].

Bloomfield et al. [32] performed experiments on rats and sheep to understand how dietary nitrate influenced thyroid glands. It was found that dietary iodine level appears higher when nitrate exists in the diet, which indirectly confirmed its association with thyroid cancer development. Warda et al. [314] agreed with the perspective that the dietary nitrate intake would affect the iodine level, leading to DTC development. On the contrary, based on the systematic analysis conducted by Bahadoran et al. [22], there was no direct association between dietary nitrate with thyroid cancer, but exposure to nitrite is somehow related. Therefore, the relationship between dietary nitrate and thyroid cancer development remains uncertain, while other studies demonstrate that environmental pollutants affect thyroid glands' functioning [160, 245, 22].

Vitamin D Deficiency

Vitamin D, as a micro-element, is sensitive to thyroid glands and is responsible for balancing calcium and phosphorus homeostasis for preserving bone health [214]. Muscogiuri et al. [214] suggested that vitamin D would increase tumour suppressor protein to prevent thyroid cancer. Zhao et al. [349] confirmed that vitamin D deficiency acts as a risk factor for thyroid cancer development. However, there are other debates around this factor. Laney et al. [153] once evaluated 24 thyroid nodules and found vitamin D deficiency rate is quite similar in both benign and malignant nodules. According to the pilot study conducted by Jonklaas et al. [131], selenium concentrations are related to higher thyroid cancer stage, while no direct associations were found between vitamin D concentration with thyroid cancer. Similarly, many scholars requested to include further investigations to identify a determinate association between vitamin D deficiency and thyroid cancer development [306, 145].

Body Mass Index, Obesity, and Diabetes

Body mass index (BMI) is considered one of the most commonly studied factors for thyroid cancer development, which is usually used to diagnose obesity. Obesity is a multi-

factorial disease developed from "interaction with genotype and environment" [84]. BMI is calculated using the $weight$ (kg) to be divided by the $height^2$ (m) [208]. More specifically, people with BMI levels higher than $30kg/m^2$ are considered obese [208]. Since thyroid cancer is the most commonly diagnosed endocrine cancer and the instances of obese people are also rising simultaneously, the hypotheses of their associations were usually established [239]. Several studies demonstrated a positive association between BMI and the existence of a higher risk of thyroid cancer, including Han et al. [101], Pappa and Alevizaki [239], Zhao et al. [347], and Zhai et al. [340]. Obesity has also been found positively related to thyroid cancer development [184, 266, 141, 116]. Besides, other risk factors were also investigated, such as excessive protein intake, carbohydrate consumption [193], and unhealthy lifestyle [245, 6]. However, the connection between diabetes and thyroid cancer development is relatively weak compared to BMI and obesity [142, 334].

Hormonal Factors

The gender disparity cause of thyroid cancer is still under-researched. Since women have strikingly higher possibilities of diagnosing thyroid cancer than males, there are always debates about hormonal factors. Horn-Ross et al. [110] proposed that factors like transient effects of pregnancy, delayed pubertal development, progesterone deficit, or estrogen unopposed by progesterone, are particularly sensitive for young females. In addition, Mannathazhathu et al. [192] applied a case-control study and confirmed that female hormones during the menstrual cycle and pregnant status are sensitive to thyroid cancer development. However, the reported findings on the gender disparity are inconsistent since many other studies failed to present a clear association between hormonal factors and the development of thyroid cancer [203, 310, 116].

Research Gaps Summary

Among the quantitative method-based pathogenesis-related literature, logistic regression (LR) and multi-variable analysis techniques were used relatively often to confirm

the identified risk factors further. The data mining techniques and the association rule mining (ARM) methods have also been applied several times. ARM is responsible for revealing hidden patterns among complex, high-dimension, and large volumes of interwoven attributes [133]. With the continuous progress of machine learning techniques, increased efficiency in revealing the hidden patterns for thyroid cancer pathogenesis has been demonstrated by the utilisation of ARM algorithms.

With the ARM techniques, the most strongly-related factor with thyroid disease is elder age. People aged from 60 to 80 are very likely to develop thyroid diseases like hypothyroidism or hyperthyroidism. Zhai et al. [340] once deployed qualitative analysis and confirmed that middle-aged people, specifically those aged from 50 to 54, are at high risk of being diagnosed with thyroid cancer compared to other age groups. Contradictory studies are also often presented. A case-control study was conducted by Mileva1 et al. [206]. From their results, it was demonstrated that younger age is statistically significantly related to the risks of establishing thyroid cancer. The retrospective study implemented by Azizi and Malchoff [21] also indicated that thyroid cancer is likely to be established in people younger than 55.

Table 2.2 presents the overall identified thyroid cancer risk factors from the past 10 years. Based on this comprehensive analysis, prior prospective studies determined that well-established factors like radiation exposure in childhood [31] and gene heredity [107] are in global consensus. In contrast, other factors like diabetes [334], obesity [99], vitamin D deficiency [26], and hormonal factors [116] are still under debate and require further investigations. Accordingly, in this thesis, the associations between these factors with thyroid disease establishment were systematically analysed.

In general, many mysteries around thyroid cancer are still yet to be revealed. One of the basis is to understand its pathogenesis by identifying the risk factors and examining their inner correlations. Although much effort has been put into the literature to investigate risk factors associated with thyroid cancer, many factors are still controversial. Therefore, in this thesis, a thorough investigation was carried out to deeply comprehend

Table 2.2: Thyroid cancer risk factors (last 10 years).

Studies	Method	Findings
[334]	Survey	Diabetes
[194]	Survey	Insulin resistance, Inflammation, Sexual hormones
[360]	Survey	Iodine deficiency
[266]	Survey	Obesity, Overweight, Radiation exposure
[134]	Survey	Radiation, Smoking, Alcohol, Nutrition
[216]	Survey	Hashimoto's thyroiditis, Elevated TSH
[340]	Survey	Geographical factors, Age, Higher BMI
[26]	Survey	Iodine deficiency, Diabetes, Pollutants, Radiation
[152]	Survey	Stress
[124]	Survey	Consumption of meat, Regular use of multivitamins, Dietary nitrate
[31]	Survey	Radiation exposure during childhood, Family history, Hashimoto's thyroiditis
[107]	Survey	Genetic factors
[310, 44]	Survey	Late age at menopause
[182]	Cohort	Hysterectomy
[325]	Case-control	Radiation, Obesity, Tallness, Artificial menopause, Family history, Iodine deficiency, Spring drinking water
[326]	Case-control	High body surface area, Great height, Excess weight, High body of fat percentage
[192]	Case-control	Hormonal factors
[181]	Multi-variable	Hashimoto's thyroiditis, Autoimmunity
[361]	LR	Smoking, History of thyroid disease, Diabetes, Radiotherapy of head/neck
[18]	LR	Marital status, Family history, Dietary iodine, Oxidative stress, Fast and fried food
[329]	LR	Obesity, Family history, Use of thyroxine
[101, 347]	Mixed	BMI, Obesity
[225, 143, 238]	Mixed	History of thyroid disease
[245, 17, 176, 76, 103]	Mixed	Radiation exposure
[239, 184, 151, 198, 14, 141]	Mixed	Overweight, Obesity

the cause of thyroid cancer development by finding correlations among the identified factors so that society can have a deterministic understanding of the epidemiology and eventually establish a consistent sense of its cause. Moreover, bridging such a gap will potentially be a breakthrough in the clinical domain.

2.3.2 Diagnosis and CAD Implementations

With the emergence of deep learning techniques, CNN incorporating medical images has been widely applied in the medical area for diagnosing breast cancer [362], heart disease [172], and liver cancer [330]. Thyroid cancer diagnoses have also been implemented universally in the deep learning area. CAD implementations for thyroid cancer detection generally consist of four components: pre-processing, segmentation, feature extraction, and classification [15].

Medical images acquired from diverse institutions are not standardised due to the utilisation of different devices, archive policies, or acquisition strategies. Therefore, pre-processing of the acquired dataset is required to remove noises, enhance image quality (i.e., contrast, colours, and sharpness), or augment the dataset to make it more adequate for model training [47]. Afterwards, the segmentation step locates the region of interest (ROI) from the background for disease detection [186]. The feature extraction step selects features from the ROI based on domain expert knowledge, allowing them to form into a feature set to be fed into the classifier for decision-making [15]. Classification is always the ultimate goal, which decides the class of the object (e.g., benign or malignant, stage of a particular disease) based on the extracted features [15]. Compared to manual diagnosis, CAD makes the diagnosis more accurate and efficient. CAD mitigates human false-positive and false-negative rates and achieves automatic diagnosis through computational power, assisting clinicians with diagnostic decision-making and allowing them to focus more on patient care.

Pre-processing

Machine learning techniques make predictions or decisions based on the learnt experience, which relies on the quality of the input information. Therefore, with the increased volume of the input data, more experience can be learnt so that the final predictions will be more accurate. In this regard, many data augmentation techniques were proposed in CAD design. For instance, Chan et al. [47] once applied horizontal flipping and contrast adjusting techniques for augmenting the original acquired $1,791$ ultrasound images to $7,360$ for thyroid cancer detection. Chouiha and Amamra [58] once adopted an open-access ultrasound image set for thyroid nodules recognition, and they have augmented the original 451 images to 4000 images. Other studies choose to crop images into patches to increase the size of the input data [348], rotate images [5], or adjust the Gaussian noises for image augmentation [276]. With varied extend of augmentation, the CAD performance can be increased, thus becoming an indispensable step for pre-processing.

Apart from data augmentation, standardising the input medical sets is also critical in the pre-processing step to reach consistency. The common tasks are removing image annotations [233] and speckle noises [300]. Chi et al. [55] deployed the artefacts removal technique proposed by Narayan et al. [218] for thyroid ultrasound images. More specifically, the authors extracted the non-zero region from the input image, plotted a histogram containing the artefacts, identified the histogram peaks as the intensity level of the artefacts, and subtracted the artefact pixels with the intensity levels to restore the image without any annotations [55]. Besides, the adaptive median filtering (AMF) technique was often used to remove annotations, markers, and noises [212]. The median filtering algorithm detects the impulse noise by comparing each pixel to its neighbours. When an impulse noise pixel is identified, its value will be replaced by the median value of all the neighbours [230]. The difference between the median filtering algorithm and the AMF algorithm is that the filter size of the latter one can be changed based on the characteristics of the input image [233]. AMF is very efficient in restoring image quality and has been used relatively often for processing thyroid ultrasound imaging [229, 226, 232].

Segmentation

The image segmentation task allows extracting ROI from the background, mitigating the use of computational resources to diagnose the overall image and enhancing the diagnostic performance by omitting insignificant features [37]. Medical image segmentation subsumes varied exceptional techniques tailored for semantic entity extraction in the computer vision domain [87]. Poudel et al. [250] once evaluated three segmentation algorithms, including active contours without edges (ACWE), graph cut, and pixel-based classifier with thyroid ultrasound images. The ACWE technique manually initialises an ROI from the input image, in which the circled region will be denoted as 1, and the rest of the ROI will be denoted as 0 as the background. Then, the ROI will be computed by using the Euclidean distance.

Another method used for thyroid image segmentation named "graph cut" was inspired by the "GrabCut" technique, which was proposed by Rother et al. [264]. The graph cut technique initially requires the manual marking of the ROI. The segmented ROI and the background will form into the Gaussian Mixture Models (GMMs) through clustering. This approach allows assigning the Gaussian components to the corresponding foreground and background pixels. Lastly, the graph cut will find the new foreground and background pixels based on the clusters.

The pixel-based classifier was also used for thyroid image segmentation [249]. It clicks inside and outside of the ROI and passes the features from both regions into the decision tree for training. In this case, the extracted features are the coefficient of variation and the mean of the two neighbourhoods. Thus, the decision tree can automatically learn the foreground and the background from those features. However, this technique requires rigour selection of ROI, otherwise, the segmentation of thyroid nodules can result in erroneous.

Besides the aforementioned techniques, CNN architectures have also been deployed relatively often for the thyroid nodule segmentation task [186, 187, 250]. With the emergence of deep learning, artificial neural networks (ANN), CNN, and recurrent neural networks

(RNN) have become backbones for image analysis, pattern recognition, and computer vision tasks. ANN and CNN are usually applied for classification tasks, whereas RNN is tailored for natural language processing for time-series predictions. The concept of CNN can be traced back to the 1980*s* [98], as it emerged from the brain's visual cortex and has been widely used for image classification and pattern recognition since then. U-Net is a well-known segmentation CNN model, which consists of down-sampling, up-sampling, and skip connection modules, aiming for biomedical image segmentation tasks [262]. The down-sampling layers of U-Net utilise convolutional operations to extract features from the input image. The up-sampling layers restore the extracted features using the down-sampled latent information. The skip connections feed the down-sampling feature maps to the corresponding up-sampling feature maps; in the meantime, crop the image from down-sampling to up-sampling to ensure the size is consistent. Besides Poudel et al. [250], many more studies deployed the U-Net architecture for thyroid nodule segmentation [354, 37, 285, 63].

With the basic U-Net structure, many more advanced models were built upon it. For instance, Ding et al. [71] modified the U-Net to embed residual and attention blocks and called ReAgU-Net. The model presented an increased dice similarity coefficient (DSC) score of 0.869 compared to the U-Net of 0.820 on the thyroid nodule segmentation task. Similarly, He et al. [106] proposed an attention-based U-Net. Nugroho et al. [231] deployed Res-U-Net proposed by Cao and Zhang [42] to segment thyroid nodules on ultrasound. Yang et al. [332] built a dual-route mirroring U-Net called "DMU-Net". Zhang et al. [343] proposed a cascade U-Net for thyroid nodule segmentation and classification. Shahroudnejad et al. [274] introduced residual dilated U-Net (resDUnet) for thyroid nodule segmentation, whereas Kumar et al. [150] deployed dilation in the same task. The dilated convolution is referred to as "convolution with a dilated filter" [337]. In particular, the dilation operation supports the exponential expansion of the receptive field without loss of resolution. The number of parameters is identical through dilated convolutions, whereas the receptive field scale grows exponentially, allowing more features to be captured.

Owning to the performance improvements in deep learning techniques, there has been a concurrent rise in researchers shifting their focus from simply CAD application to designing more exquisite and sophisticated models or modules for the thyroid nodule segmentation task. Li et al. [165] applied a fully convolutional network (FCN), in which the model only performs convolution, pooling, and up-sampling. Likewise, Gong et al. [92] proposed an encoder and decoder-based FCN model for thyroid nodule segmentation and achieved 81.26% DSC outperforming U-Net. Zhou et al. [353] also deployed the encoder-decoder structure on thyroid ultrasound images. Those segmentation-related studies harnessed the power of deep learning algorithms dedicated to detecting thyroid ROI more efficiently and precisely, leading to enhanced classification accuracy.

Feature Extraction and Selection

The features extracted manually from medical images are called "hand-crafted features". Most existing studies consider hand-crafted geometric, morphological, and texture features from ultrasound images in the CAD setting for thyroid cancer diagnosis. Geometric features are the information used to evaluate an object with geometric elements to describe the shape of irregularity [91]. Morphological features are information about lesions' morphological characteristics [333]. Additionally, texture features are represented by an image's contrast [333]. Gomes Ataide et al. [91] extracted geometric and morphological features from the open-access Digital Database of Thyroid Ultrasound Images (DDTI). By augmenting the image set, the authors have extracted 19 geometric (e.g., solidity, rectangularity, Orientation, roundness, centroid, etc.) and 8 morphological features (e.g., area, perimeter, area perimeter ratio, etc.) from $3,188$ ultrasound images. With the extracted features, the classifier reached an accuracy of 99.33% for detecting malignant thyroid nodules. Similarly, Nugroho et al. [230] deployed 9 geometric features (e.g., circularity, compactness, convexity, solidity, etc.) from 165 ultrasound images. Through standardised pre-processing, segmentation of ROI, and feature extraction steps, the neural network reached an accuracy of 0.9479 classification performance. Yang et al. [333]

combined textual and morphological features (e.g., solidity, centroid, bounding box, etc.) for thyroid cancer detection, and their work obtained a diagnostic accuracy of 89.13%. From fair to moderate agreements, it has been reported that hand-crafted features are beneficial for establishing thyroid cancer detection CAD systems.

Apart from ultrasound images, a few studies proposed other image modalities for feature extraction in thyroid cancer detection. For instance, Wei et al. [316] and Hu et al. [111] both applied radiomics analysis on magnetic reasoning images (MRI). Lu et al. [180] and Zhou et al. [355] adopted computed tomography (CT) scans for radiomics analysis. Wu et al. [324] adopted morphological features from CT scans, and their work reached a classification accuracy of 77.7% for detecting PTC.

Based on the identified studies, researchers tend to apply various feature selection algorithms to establish thyroid cancer diagnostic systems. The commonly selected features from ultrasound images for CAD design are summarised in Table 2.3. The most commonly used features are nodular size, shape, margin, composition, and calcification presence. The knowledge extracted from medical images will then be fed into machine learning or deep learning classifiers to evaluate the thyroid status for disease detection.

Classification

With the astonishing development of electronic computers, three basic machine learning algorithms emerged in the $1950s$, including symbolic learning, statistical learning, and neural networks [148]. Those three branches grew more advanced and become the well-known classifiers today "decision trees (DT), K-nearest neighbours (KNN), and multi-layer feed-forward neural networks" [148]. Various CAD systems were built upon those classifiers, such as support vector machines (SVM), Naïve Bayes (NB), and multi-layer perceptron (MLP), to name a few. Deep learning emerged as a subset of machine learning and has become an intense tool for computer vision tasks. It allows the automatic classification of the extracted features [29]. Recently, deep neural networks have been used frequently in CAD design for helping to make diagnostic decisions and have shown

Table 2.3: Ultrasound hand-crafted features for thyroid cancer detection.

Reference	Features Used
[168]	Intercept, size, shape, margin, echogenicity, cystic change, micro-calcification, halo
[177]	Micro-calcification, shape, margin, capsular invasion, architecture, echogenicity, ring down artifact, vascularity
[49]	Gray level co-occurrence matrix, statistical feature matrix, gray level run-length matrix, law's texture energy measure, neighboring gray level dependence matrix, wavelet features, Fourier features
[357]	Shape, margin, echogenicity, internal composition, presence of calcification, peripheral halo, vascularity
[3]	Shape, echogenicity, calcification, echo texture, margin, capsule invasion, halo
[135]	Size, area, shape, color, texture of regions, histogram of oriented gradients, co-occurrence gray level matrices, chromatin density
[50]	Intensity differences, elliptical fit, gray-level co-occurrence matrix, gray-level run-length matrix
[322]	Location, position, shape, margin, internal contents, echogenicity, calcification, echo-genic foci in solid portion, halo, infiltration and extracapsular invasion, increased intra-nodular vascularity, abnormal lymphadenopathy, multifocal
[173]	Gray level co-occurrence matrix, Local binary patterns, Histogram of oriented gradient, Scale-invariant feature transform, vector of locally aggregated descriptors
[146]	Margin, internal content, anteroposterior dimension-to-transverse dimension ratio, microcalcifications
[174]	Size, margin, shape, aspect ratio, composition, calcification
[166]	Size, morphology, location, echo, margin, boundary, surrounding tissue, posterior echo, calcification
[308]	Aggressive histology subtype, vascular tumour capsular invasion, extra-thyroidal extension, metastases
[90]	Composition, echogenicity, orientation, margin, shape, spongiform, calcification, elasticity, vascularity
[335, 122]	Composition, shape, margin, orientation, echogenicity, spongiform
[45, 95, 278]	Composition, echogenicity, calcification, margin, shape
[327, 234, 341]	Size, margins, shape, aspect ratio, capsule, hypo-echoic halo, internal composition, echogenicity, calcification pattern, vascularity, and cervical lymph node status

satisfying performance, such as detecting diabetic retinopathy [338], Covid-19 detection [309, 16], and malaria diagnosis [167]. Studies using machine learning-based CAD application to detect thyroid cancer is also abundant.

Liu et al. [175] once adopted Naïve Bayes (NB) on 41 thyroid nodules (21 benign and 20 malignant) from 37 patients and compared with two experienced radiologists. The results obtained by the NB algorithm outperformed the radiologists and reached an area under the curve (AUC) of 0.851. Similarly, Singh and Jindal [282], Xia et al. [327], and Ouyang et al. [234] acquired self-obtained ultrasound features for making classifications and obtained comparable results to experienced radiologists.

In some earlier works, ANN has been adopted relatively often by using ultrasound images to make a thyroid cancer diagnosis. According to Zhu et al. [357], 689 thyroid nodules were examined using ANN and reached a classification accuracy of 83.1%. Shin et al. [278] compared the SVM classifier with ANN through 348 thyroid nodules, and the diagnostic accuracy rates were 69% and 74%, respectively. More recently, CNN has been adopted more frequently for detecting thyroid cancer.

CNN models play vital roles in the computer vision domain since they can be used for segmentation, localisation, and classification tasks [37, 20, 336]. Li et al. [164] once acquired 131,731 ultrasound images for cross-institutional analysis through the CNN model on the thyroid cancer diagnosis. Their work demonstrated an accuracy of 0.889, 0.856, and 0.915 for the three cohorts from China, including Tianjin, Jilin, and Weihai. Buda et al. [36] acquired ultrasound images from an institution in Germany and applied CNN for thyroid cancer detection. Their work obtained an accuracy of 0.78, which was lower than the respective results of experienced radiologists. Zhu et al. [359] proposed a generic eight-layer CNN model for classifying thyroid and breast lesions. With the thyroid cancer detection task, the CNN model reached an accuracy rate of 86.5% on 719 ultrasound images. Nugroho and Frannita [227] used the Inception model to detect thyroid cancer and reached an accuracy of 87.2% with the DDTI ultrasound repository. Chan et al. [47] once compared VGG19, ResNet101, and InceptionV3 models on 812 ultrasound images, and

the best-performing model was ResNet with a 77.6% accuracy rate reached for thyroid cancer diagnosis.

The use of CNN models demonstrates varied performance following the different image analysis steps with heterogeneous image quality. In most cases, the proposed models are incompetent in generalising to different datasets, and there is also a possible over-fitting concern for most existing works. In this regard, transfer learning has been applied quite often to mitigate the over-fitting phenomenon. Transfer learning uses parameters learnt from pre-trained neural networks and applies those "gained knowledge" to new tasks by freezing the previous layers and making changes to the last few layers. Ma et al. [188] proposed the integration of two pre-trained CNNs through transfer learning, where the shallower network was used for learning high-level abstract features, and the other deeper network was used to learn low-level detail features. Then, the two learned feature maps from two CNNs were fused as an input into a softmax layer to diagnose malignant thyroid nodules, resulting in diagnostic accuracy of 83.02%. Chi et al. [55] fine-tuned an Inception model and tested it on two ultrasound databases. For the DDTI set, the accuracy was 98.29%, and for the private self-acquired data set, the accuracy was 96.34%. A moderate consensus was made that CNN applications on ultrasound images for detecting thyroid cancer are efficient and accurate [146, 95, 166, 5].

There has been a concurrent rise in applying other image modalities for thyroid cancer detection rather than ultrasound. For example, Bakht et al. [23] deployed transfer learning on AlexNet and VGG models with cytology images and reached an accuracy of 93.05% obtained by VGG19. Wang et al. [313] adopted VGG19 and InceptionResNetV2 to multi-classify thyroid nodules into seven classes through histopathology images, including normal tissues, adenoma, goitre, papillary cancerous nodule, follicular cancerous nodule, medullary cancerous nodule, and anaplastic cancerous nodule. The results suggest that VGG19 yields better averaged accuracy for the seven classes than InceptionResNetV2, which is 97.34% and 94.42%, respectively. Similarly, Dov et al. [72] adopted the Multiple-instance Learning (MIL) approach on segmented whole-slide images to predict the malig-

nancy of thyroid tissues. Buddhavarapu and J [38] trained ResNet50 and DenseNet121 models through transfer learning with histopathology images and reached an accuracy of 100%. Chandio et al. [48] acquired cytological images to detect medullary thyroid cancer, and the CNN model reached an accuracy of 99.00%. Apart from those, Lee et al. [158] adopted eight CNNs on CT scans to differentiate thyroid cancerous metastasis, including DenseNet121, DenseNet169, InceptionResNetV2, InceptionV3, ResNet, VGG16, VGG19, Xception. Based on their comparison, the best AUC was obtained by InceptionV3 and ResNet, which was around 0.95. Zhang et al. [342] utilised MRI on multi-modality CNN for thyroid disease classification and reached a diagnostic accuracy of 0.82 on 45 images. Moreover, Naglah et al. [217] constructed a multi-input CNN for thyroid cancer diagnosis from 49 patients who underwent MRI tests, and their model reached an accuracy of 0.88. Based on investigations, ultrasound seems to lose its dominance in thyroid cancer detection for CAD design since more medical image modalities were analysed through the use of CNN, including but not limited to CT scans [345], hyperspectral imaging [100], and SPECT images [189].

Research Gaps Summary

Thyroid disease diagnosis is correlated with precision treatments to achieve a promising prognosis. Thyroid disease has several types, all resulting in different treatment protocols. For the sake of customised treatments and a well-established prognosis, the diagnosis should be targeted more. However, the existing studies mainly focus on the binary classification to detect hypothyroidism/hyperthyroidism or benign/malignant thyroid nodules. Among all the identified thyroid cancer CAD-related works in the literature, only one study has applied the multi-class classification task to determine thyroid cancer subtypes [313]. However, sound results were demonstrated for classifying the types after surgery. Under this situation, pathologists can experience reduced workloads, whereas patients cannot benefit much. Therefore, studies around the multi-class classification of thyroid cancer are significantly limited and should be further propagated.

Furthermore, the existing works have considerably ignored the thyroid disease subtype co-existence phenomenon. Most existing studies are dedicated to applying ultrasound images for making diagnoses for individual nodules. Relying on the diagnostic decisions made for each nodule at a time is inefficient. More importantly, the existing CAD models usually are applicable to unitary datasets and are incompetent in adapting to different institutions. Therefore, providing professional human-level diagnoses for thyroid disease patients is highly encouraged, and it is expected to elevate the clinical adoption of CAD models by improving their generalisation.

2.3.3 Prognosis and Recommendations

With the progressive development of CAD systems, a manifesto for promoting an accurate diagnosis has been achieved, whereas precision medicine is a prospective phenomenon to be launched. Individual patients have varied health conditions or disease stages, while treatment generalisation is inappropriate. Precision medicine, on the other hand, is tailored for individual health care on the basis of the target's genes, lifestyle and environment [109], maximising the patient's health status after targeted treatments. In the clinical domain, customised treatment plans contrive to improve individual prognosis. On top of that, the establishment of sagacious decision support systems on treatment protocols enlarges the applications of precision medicine. More importantly, predicting thyroid cancer patients' death, survival, and recurrent rates will potentially guide them to achieve an optimal prognosis.

Treatment Decision Support System

Thyroid disease treatment protocols heavily rely on individual health conditions, such as age, weight, BMI, pregnancy status, medical history, comorbidity, and medication dosages. Various external factors like seasonal temperature, financial status, and the patient's preference also play significant roles in developing treatment protocols [272]. In this regard, precision medicine and treatment must be achieved for the target patient.

With the increasing instances of thyroid cancer, CAD applications have been generated more often, whereas precision medicine development has been somehow neglected in the past few years. Among the identified studies, the development of customised treatment decision support systems-based works is quite limited. Chen et al. [52] once adopted the density-peaked clustering analysis technique for disease symptoms clustering. Meanwhile, the authors adopted ARM for establishing treatment rules, called the Disease Diagnosis and Treatment Recommendation System. The ARM algorithms were used to detect the associations between the symptom clusters and the treatment schemes, and the system yielded a satisfying performance. Meanwhile, the system's interface was also designed and tested.

Katzman et al. [132] once applied the DeepSurv model to develop personalised treatment protocols for patients with a particular disease. The DeepSurv is a feed-forward "Cox proportional hazards deep neural network" that is used to model the interactions between a patient's covariates and treatment effectiveness [132]. Besides the DeepSurv models, the fuzzy cognitive map (FCM) approach has also been implemented a few times to design a patient-centric treatment decision support system. An FCM incorporates ANN and fuzzy logic and shares similar logic with human reasoning and decision-making [237]. More specifically, FCM integrates qualitative and quantitative data, and it looks like a cognitive map consisting of concepts and relationships. In addition, FCM can model complex systems and is tailored for developing decision-making systems, particularly disease treatment decisions. FCM consists of concepts and weights. The concepts are the representative variables for making treatment decisions, including patient age, nodule size, nodule location, and metastasis extent, to name a few. The directed edges with arrows present the degree of the relationship between the interdependent concepts, known as weights. Moreover, the FCM approach has been applied several times in designing personalised disease treatment recommendation systems [236, 237, 288].

Death, Survival, and Recurrence Prediction

Several judicious studies proposed the utilisation of machine learning approaches to evaluate the prognosis of thyroid cancer, including the survival rate, death rate, and recurrence rate of patients diagnosed with thyroid cancer. Among the identified papers, ANN has been adopted most often [61]. Jajroudi et al. [119] have applied the SEER dataset on Multi-layer Perceptron (MLP) and LR to predict the survival rate for 7,706 patients with thyroid cancer, resulting in a better performance obtained by the MLP approach. Mourad et al. [211] has applied three MLPs to the SEER dataset to determine the probability of death rate caused by thyroid cancer. Their first MLP has selected input features, including patient gender, age, race, tumour size, primary disease extent, nodular location, and the number of positive lymph nodes. With a 19-layer MLP, the ANN would output the patient's status, which is still alive or dead. The second MLP only inputs age, primary disease extends, and nodule locations selected by the filter-based feature selection algorithms with 18 hidden neurons and the same output. The third MLP is designed with three different input features, namely tumour size, number of positive lymph nodes, and metastases. Moreover, the architecture has 4 hidden neurons, and the outputs are maintained from the second MLP architecture. The best performing model was the first MLP that reached an accuracy of around 94.49%.

Researchers also pay close attention to the recurrence rate of thyroid cancer. For instance, Park and Lee [240] compared five different machine learning models to evaluate the recurrence rate of patients with PTC. The selected input features were age, gender, tumour size, multiplicity, lymph node metastasis, lymph node ratio, extranodal spread, and extrathyroidal extension. The selected models were DT, RF, XGBoost, LightGBM, and Stacking. The DT outperforms the other four algorithms and has reached a prediction rate of 95%. Yang et al. [331] applied an Ensemble Approach for Clustering Cancer Data (EACCD) to design a prognosis system for thyroid cancer patients to minimise the probabilities of recurrence. Through a set of input features, such as tumour size, lymph nodes, metastasis, and age, the EACCD algorithm was applied. It consists of

the following three steps: defining the initial dissimilarities between classes, applying ensemble learning to obtain the learned dissimilarities, and clustering the combinations of the learned dissimilarities. Schneider et al. [270] applied the multivariate LR on 217 patients' lymph ratios to determine the recurrence of papillary thyroid cancer. The results show that the lymph node ratio is a significant factor that correlates with PTC recurrence.

Research Gaps Summary

Patients diagnosed with thyroid cancer concerns the most about the treatments, recurrence, survival, and death rates. Based on our analysis, studies on machine learning-based treatment decision support systems are far less than those on diagnostic systems. Among all the identified works, none of them was designed and targeted specifically for thyroid cancer. Unlike the CAD systems, the treatment decision support systems are much more interpretable and acceptable by clinicians since factors affecting treatment decision-making are easily comprehended and quantified. Moreover, clinicians are crucially involved in the system's development phase. However, the works around this area have been considerably ignored. With the customised treatment decision support system design, individual patients can obtain optimal prognostications by receiving the most proper treatment protocols so that the recurrence or even death rates can be dramatically reduced.

Today is the artificial intelligence era, and many more advanced machine learning techniques have emerged yet have not been applied in the clinical field for thyroid cancer treatment. Hence, a thyroid disease-specific treatment decision support system should be generated by adopting current state-of-the-art algorithms. Such a system should consider a comprehensive list of factors to fill in the gaps presented in clinical treatment guidelines, such as information around demographic, medical history, dietary, financial conditions, disease subtypes, illness degree, comorbidity, nodular characteristics, and medication intake, to name a few. Through comprehensive factor analysis, an optimal treatment plan can be provided for patients to achieve customised treatments to enhance clinical trust

leading to potential adoption.

Another worth-mentioning point is that the time factor is crucial for the survival, death, and recurrent rate prediction task. However, existing studies have significantly neglected this factor during the model design phase. It is highly encouraged to incorporate the time factor into building a sophisticated model to forecast predictions to help patients prepare for immediate health condition changes. These limitations will all be re-mentioned and addressed in Chapter 8 to be deployed in future work.

2.4 Summary

This thesis addresses thyroid cancer pathogenesis, diagnosis, and prognosis challenges. Following a rigorous procedure of the proposed literature review framework, this thesis involves multi-disciplinary studies, including clinics, statistical analysis, computer vision, and machine learning. Based on the comprehensive investigations, the literature gaps and the corresponding research questions are summarised as follows:

- **RQ** 1*: What pathogenesis of thyroid disease can be ascertained through data mining approaches?*

  The research in the field of identifying thyroid cancer-related factors is well established. However, many risk factors were derived from meta-analysis without quantitative investigations. Some factors were identified with insufficient scales or sources of datasets, making thyroid disease's pathogenesis controversial and requiring further determinations through data mining applications.

- **RQ** 2*: How can the diagnosis of thyroid disease be improved via deep learning techniques?*

  Most identified studies built CAD systems for thyroid disease detection, including distinguishing between hypothyroidism and hyperthyroidism and identifying malignant nodules from benign ones. Nevertheless, the existing studies on differentiating

among thyroid disease types are significantly limited. Moreover, the disease subtype co-existence phenomenon has also been ignored by existing studies. Therefore, this thesis sought to provide expert-level diagnostic decisions automatically through deep learning techniques to mitigate human false-positive and false-negative diagnostic rates.

- **RQ** 3*: How can a deep learning-based model be adaptive to different patient profiles for thyroid cancer diagnosis?*

  The existing CAD systems usually utilise medical sets retrieved from unitary data sources. Such systems cannot be generalised to different data sources with distinct patient profiles, resulting in the varied diagnostic performance for the same model under distinct datasets. Hence, a unified model that can be adaptive to different patient profiles will potentially enhance clinical trust and implementation, essentially helping elevate patient prognosis.

Chapter 3

Research Paradigm and Methodology

A research paradigm articulates a pattern of understandings from the theory or belief of the research project which was built, and it usually exists in the form of a philosophical framework [287]. A research paradigm normally reflects the perception towards the essence of the reality under the study. It usually holds assumptions around ontological, epistemological, and methodological concerns [127].

Ontology refers to the beliefs about the nature of reality [113]. Researchers tend to inquire about reality, including: How does it exist? What can be known from it? Is reality subject to perceptions? Those concerns challenge researchers to question the beliefs about reality and are known as ontological concerns [259].

Epistemology describes the beliefs about knowledge, findings, and relationships between researchers and studies [113]. The epistemological questions usually debate and incorporate "objectivity, subjectivity, causality, validity, and generalisability" for acquiring knowledge [241].

Methodology spans from broad assumptions to the detailed methodological process [113]. It usually articulates how research should be undertaken, including data collection and analysis of data production [259]. It involves the strategy, planning, and processing of research [60]. The methodological concerns should be around how the research should be conducted and how to validate the findings.

In short, ontology questions the nature of the world, epistemology inquires about the kind of knowledge acquired from the world, and methodology decides on the procedures for obtaining the knowledge [127]. In this thesis, two types of research paradigms were considered and adopted, namely *positivism* and *interpretivism*.

The *positivism* paradigm comes in as objective to observe a phenomenon without interference [259]. Positivist methodology tends to investigate the existence of an association among variables rather than interpreting the findings. It usually generates and relies heavily on quantitative data, where a hypothesis is put forward, evaluated based on statistical experiments, and analysed with empirical evidence [67].

The *interpretivism* paradigm is relatively more subjective than *positivism* [259]. The goal is to comprehend the phenomenon, knowledge, or reality from the context where the individuals interact with each other [259]. Interpretive researchers mainly generate, employ, and analyse qualitative data. The use of numerical data in interpretive studies is limited and not critical.

However, the *positivist* and *interpretive* paradigms were criticised largely by scholars [259]. Specifically, the *positivism* methods were criticised for being appropriate when analysing the natural phenomenon, whereas they cannot fulfil the demand for interpreting social phenomenon research [86]. The *interpretivism* methods, on the other hand, have limitations in ensuring the credibility, transferability, reliability, and objectivity of the findings [96]. Therefore, integrating the two research paradigms can essentially form a more robust strategy for research purposes as they complement each other and maximise the validity of the research study.

3.1 Research Paradigm

This thesis follows a mixed methods research paradigm involving qualitative (i.e., *interpretivism*) and quantitative (i.e., *positivism*) analysis [130]. Qualitative purists highlight the superiority of constructivism and interpretivism; quantitative purists main-

tain objectivity in social science [130]. Both sets of the analysis view their research methods as ideal, impeccable, and implicitly advocate incompatibility [130]. While the goal of the mixed methods research paradigm in this thesis is to draw from the strengths and minimise the weaknesses of both sets in multi-disciplinary studies [130].

Multi-disciplinary studies are a profusion of projects drawing together information from various domains [263]. This thesis integrates knowledge from social science, health, and data science fields and works discretely to recommend solutions to clinical challenges, considered a multi-disciplinary research study. With the establishment of CAD systems in the medical science domain, such a study will potentially benefit from multi-disciplinary collaborations, inherently elevating the reliability of decision-making [196].

The use of the mixed methods research paradigm in multi-disciplinary studies is highly encouraged [298]. Qualitative methods are dedicated to providing an in-depth and interpretive understanding of a social challenge, and the collected data is detailed and informative, whereas quantitative methods cannot reach [210]. Moreover, qualitative methods involve exploratory analysis, which generally does not have explicit hypotheses to be evaluated. Instead, the qualitative research methodologies usually ensure the research study is grounded in the researchers' and big data experiences [210].

On the other hand, quantitative methods fall under empirical and statistical study categories, and they seek to take a sampled population to investigate a pattern of a particular phenomenon [298]. Additionally, quantitative methods usually begin with a hypothesis and test for confirmation or divarication of that hypothesis [219]. It is interesting to notice that quantitative methods in the social science and medical science domains prevail more than qualitative methods, as they generally establish experimental designs to provide empirical evidence [219].

Integrating the qualitative and quantitative methods will essentially enhance the research credibility of multi-disciplinary studies [298]. Research credibility strongly relies on the research methods, including the logic of the methodological design, the data collection approach, the analytic techniques, and the evaluation criteria. Those factors are account-

able for the reliability of a research design. More specifically, the qualitative methods focus on small samples yet in-depth analysis for investigations. The quantitative methods depend upon probability samples from the original population [298]. With the integration of both methods, the mixed methods research paradigm is likely to maximise opportunities to present legitimate research findings for a multi-disciplinary study, which aims to accomplish the benefits as follows:

- Analyse divergent findings from qualitative and quantitative methods based on different methodological designs.
- Establish a consensus statement for a certain challenge from both types of analyses in the multi-disciplinary study.
- Explore the impact of the different methods on the same task and suggest addressing the challenge with a proper trade-off.

3.2 Research Methodology

Qualitative data analysis makes sense of data and transforms information into knowledge and findings [318]. The qualitative data analysis process can be defined as a "bottom-up" approach where the data is categorised, explored, explained, and mapped until patterns emerge [318]. Qualitative content analysis, as a subset of qualitative analysis, is considered a method that categorises contents based on themes and contexts [318]. The root of the qualitative content analysis is to measure the frequency of words in the context, which reside in the quantitative approach [318]. Berelson [27] once defined the qualitative content analysis approach as "a manifest of communication for the objective, systematic, and quantitative descriptions". It is regarded as an interpretative approach, which describes the themes that are informative to the research objectives [318].

Quantitative data analysis usually deals with numerical or numbers-convertible data, and their findings are generally more explicit than qualitative analysis [318]. Statisti-

cal learning is the essence of quantitative methods, which organises, analyses, interprets, and presents numerical variables to evaluate pre-defined hypotheses. The basic process of quantitative data analysis techniques involves data collection, data pre-processing, experimental analysis, and result interpretations. However, the biggest challenge of quantitative data analysis is that the analysis might be performed without understanding the appropriate techniques to be applied [318]; this is when exploratory data analysis (EDA) takes place to examine and understand the data [147].

3.3 Research Procedure

By integrating the qualitative analysis and quantitative analysis techniques, this thesis applied the mixed methods research paradigm, and the step-wise research procedure is summarised as follows:

1. ***Comprehensive literature review analysis.***

   The qualitative content analysis procedure was followed to identify thyroid disease-related risk factors based on their frequency from existing works' findings. The generated hypotheses of risk factors waiting to be evaluated were selected based on the controversial research findings from the existing works in the literature. Besides, the understandings of existing literature studies from different research objectives were evaluated independently to identify the research gaps. The investigations of the existing works were critical for the development of CAD systems in this thesis.

2. ***Data collection.***

   With the ethics approval obtained from Monash University (Project ID: 24704, in Appendix A), the data collection phase took place, and this phase accumulates multiple sources for investigations. The self-acquired datasets were obtained from a first-class hospital in China and a well-known hospital in Australia. This thesis also involved the utilisation of digital data sources, including the UC Irvine Machine Learning Repository (UCI) [73] and the Digital Database of Thyroid Ultrasound

Images (DDTI) [243]. During the datasets acquisition stage, the distribution of the acquired data samples was ensured to align with the distribution of the original population. More specifically, the gender groups, age groups, and target class numbers all maintained the distributions from the original population.

3. ***Data pre-processing.***

   This thesis analyses quantitative data from multiple source domains, and the pre-processing of those data ensures consistency, including data cleaning, data transformation, and data integration [318]. The data cleaning phase removes irrelevant data, duplicates, and outliers, whereas the missing variables are also replaced. Data transformation and integration make data in a consistent format that can be used for further investigations, such as transforming images into matrices, converting categorical variables into numerical values, and discretising data based on stratification. Data from multiple sources will need to be integrated to allow extensive analysis, and the prerequisite is to have a consistent format.

4. ***Experimental analysis.***

   Statistical analysis can be categorised into *descriptive* and $inferential$ kinds, where the former uses techniques to describe the data and the latter analyses the data to draw inferences [318]. This thesis exhibits the hybrid method, which combines both types of statistical techniques. The *descriptive* analysis was achieved with EDA methods to examine the data regarding its distribution, anomalies, dispersion, and relationships among attributes [318]. The $inferential$ analysis observed patterns and associations among variables, such as feature selection and association rule mining techniques were applied to test the hypotheses. Besides, the experimental designs for each research gap were independent (source code available in Appendix B), while the use of $inferential$ analysis is thoroughly proposed to generate the corresponding decisions.

5. ***Result interpretations.***

   The evaluations of the experimental designs were achieved with metrics, such as the area under the curve (AUC), accuracy (ACC), precision (positive predictive value, PPV), recall (sensitivity), specificity, F1, negative predictive value (NPV), and false positive rates (FPR). The overall thesis consists of classification tasks, where the class labels were binary and multi-class. This thesis addresses existing clinical challenges with respect to thyroid disease pathogenesis and diagnosis to achieve a promising prognosis. The proposed models were compared to baseline works, and the current state-of-the-art performance was achieved under several scenarios. Besides the interpretations of the generated results, the discussions around the cause of the disease, the explanations of the findings, the contributions, and future clinical implementations of the models were all presented.

By summarising the aforementioned methods, the use of the mixed methods research paradigm ensures the reliability of the generated results, and the detailed research methodology framework is presented in Figure 3.1. This thesis follows the proposed methodology framework for the implementation of the research study.

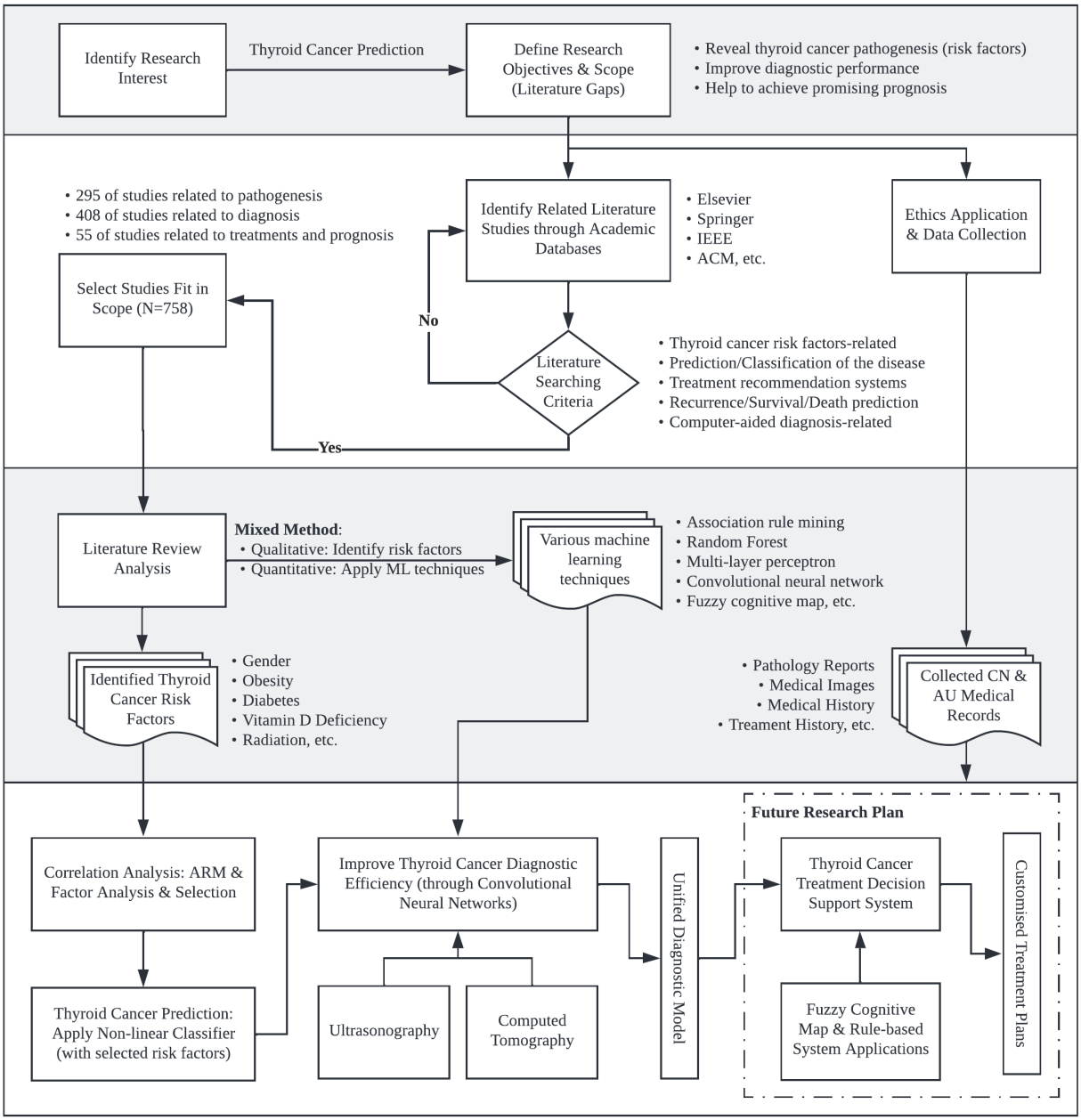


Figure 3.1: Research methodology framework.

Chapter 4

Data Mining in Thyroid Disease Pathogenesis Identification

4.1 Introduction

Thyroid disease instances are rapidly increasing worldwide, and thyroid cancer is even ranked as the fifth most commonly diagnosed disease among females in the United States [41]. Understanding the pathogenesis of thyroid disease will reduce morbidity rates, whereas thyroid cancer instances can also be lessened by avoiding the potential risk factors, leading to mitigated mortality rates. Currently, the scientific community is struggling to determine the leading cause of thyroid cancer. A few studies adopted qualitative approaches to identify potential risk factors correlated with the disease. The identified factors include radiation, depression, obesity, iodine intake, diabetes, hormonal factors, and gene heredity, to name a few [246, 83]. Nevertheless, many of those factors are still under debate and cannot be verified solely by relying on qualitative investigations.

Medical datasets are complex as multiple attributes interweave one another, making the identification of thyroid cancer pathogenesis even more challenging [235, 185, 319]. However, the investigations of the risk factors through quantitative techniques in the existing studies were analysed independently. By this approach, the interconnections

among factors were considerably ignored, thus aggravating the limited reliability of the identified factors.

Association rule mining (ARM) is a data mining technique dedicated to finding and describing relationships or hidden patterns among multiple attributes given a database [35]. In the past few decades, ARM has been applied several times in the medical domain for identifying the underlying correlations among different types of diseases [159, 46, 295, 120, 183]. Medical records are high-dimensional and complex. As a result, mining from such heterogeneous data requires significant effort and time to dig into latent patterns manually, which might be unrealistic. ARM is quite efficient for dealing with complex and sensitive data, making it appropriate for discovering unrevealed information from medical datasets. In addition, ARM, as a quantitative analysis technique, has been scarcely applied for thyroid disease pathogenesis identification. The goal of this thesis is to bridge this literature gap and assess the risk factors through ARM to ascertain their associations with thyroid disease.

The underlying patterns derived from ARM can be categorised into common rules (i.e., rules with high support and high confidence), reference rules (i.e., rules with low support and low confidence), and exception rules (i.e., rules with low support and high confidence) [296]. Common rules describe explicit information which interprets the regularity of objects with consequences; reference rules are outliers that are less meaningful and are generally excluded; exception rules, on the other hand, outline the unexpectedness of associations and are often tied up with actionability [291]. The existing studies mainly focus on adopting ARM for common rules extraction, whereas exception rules extraction has been considerably neglected [296]. Nevertheless, exception rules are potentially more engaging and valuable than common rules [170], and they can provide information that reveals unusual and contradictory but significantly meaningful knowledge.

Therefore, this chapter incorporates text mining procedures with ARM algorithms to identify and assess risk factors correlated with thyroid disease. More specifically, text mining procedures are used to identify critical features from raw medical data, including

medical history, comorbidity, and risk factors. ARM algorithms then extract common and exception rules simultaneously when incorporating an exceptionality measure, leading to valuable knowledge discovery. The identified risk factors will be compared to a set of feature selection algorithms for further determinations. The related background works of this chapter were described in Chapter 2.3.1.

The contents of this chapter have been published in journal article** 4**.

4.2 Hypotheses Formulation

Based on the comprehensive literature review analysis, the confirmed risk factors for thyroid cancer include radiation and genetic factors [82, 103], whereas the remaining factors are still controversial, such as vitamin D deficiency, obesity, iodine deficiency, dietary nitrate intake, diabetes, and hormonal factors [214, 40]. We have to underline that iodine and nitrate intake levels were usually not recorded in admission reports and were hard to be measured for patients and thus were excluded from this research study.

Given a medical data repository $\mathcal{D} \in \{X, y\}$. $X \in \mathcal{R}_i^n$ where $\mathcal{R}$ is the feature sets, such as patient medical records, including their blood examinations and risk factors, i is the ith number of patients, and n is the total number of instances given in the database. $y \in \{0, 1\}$, which denotes the classes for the instances. More specifically, this chapter extracts risk factors for thyroid diseases, including thyroid disorder (i.e., hypothyroidism, hyperthyroidism) and thyroid cancer independently. Therefore, 0 indicates that the patient is healthy (i.e., without thyroid-related diseases), and 1 indicates that the patient is sick (i.e., with hypothyroidism/hyperthyroidism or thyroid cancer).

With the qualitative analysis from the 295 pathogenesis-related studies, the controversial risk factors were hypothesised that require further data mining investigations. Consider each factor as an attribute from $\mathcal{R}$ where $\mathcal{R} \in \{a, b, c, \ldots, m\}$, which denotes the entire set of attributes to be evaluated. In this case, the factors which need determination for positively associated with thyroid disease are vitamin D deficiency, diabetes,

obesity, hypertension, and depression extracted from patient medical history reports.

4.3 Methodology

In order to confirm the correlations between the risk factors with thyroid disease development, a rigorous procedure has been followed, starting from the initial extraction of the critical attributes with text mining, correlation analysis, ARM, and feature selection investigations until the classification determinations.

4.3.1 The Proposed TM-DM Framework

Figure 4.1 demonstrates the proposed TM-DM framework which consists of three phases, including 1) text mining-based risk factor extraction, 2) correlation analysis and feature selection, and 3) classification evaluation.

Following the comprehensive literature review analysis, the risk factors have been identified and involved in the hypotheses formulation stage. In this case, the text mining procedures were integrated to validate those factors' existences from the acquired medical datasets. Specifically, the raw admission reports and discharge summaries were used for extracting critical factors. The admission reports contain the patient's demographic information like age, gender, ethnicity, medical history, lifestyle behaviours, and current symptoms. The discharge summaries include the patient's disease stage, treatment protocols, comorbidities, and principal diagnosis. After breaking the sentences into tokens, only the medical terminologies were extracted and normalised through stemming and lemmatisation, and a set of stop-words were defined and removed. The normalised attributes confirmed the presence of the hypothetical factors for individual patients denoted as 1; otherwise, the absences of the factors were denoted as 0.

The data pre-processing strategy initially transforms the format of all attributes by converting the categorical variables into numerical values. The correlation analysis was then conducted among the attributes to exclude the features with the most negligible

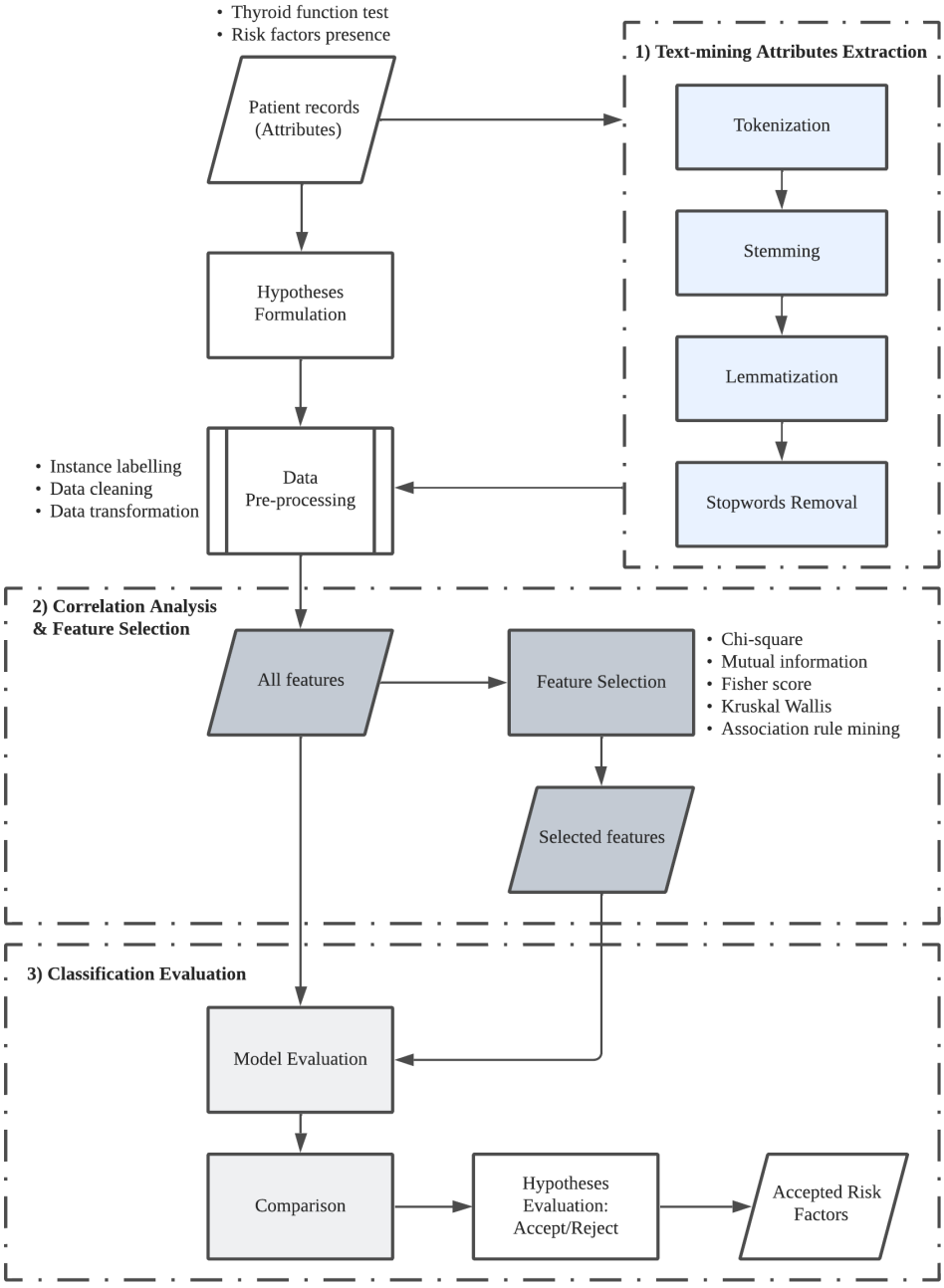


Figure 4.1: Text mining - Data mining (TM-DM) framework for thyroid disease pathogenesis identification.

impact on the class label. Feature selection and association rule mining techniques were applied to validate the risk factors. The incorporated feature selection algorithms include Chi-Squared Test (Chi) [248], Mutual Information (MI) [88], Fisher Score (Fisher) [94], and Kruskal-Wallis (KW) Test [200]. The involved ARM algorithms are Apriori [7], Frequent Pattern-Growth Tree [102], and the developed Faster Apriori algorithm. The features from feature selection algorithms were then compared to the ones derived from the ARM algorithms.

Afterwards, the selected attributes from the feature selection and the ARM algorithms were involved in a classification task for evaluation. In this case, a set of basic classifiers were implemented for evaluation, including Logistic Regression (LR), Decision Tree (DT), Support Vector Machines (SVM), Random Forest (RF), Naïve Bayes (NB), and Multi-layer Perceptron (MLP). The classification performance comparison was made between baseline models (i.e., without feature selection algorithms) and models with feature selection algorithms through the 10-fold cross-validation (CV) technique. The risk factors were determined when the feature selection algorithms proposed competitive or even outperforming results, and the hypotheses would then be rejected or accepted accordingly.

4.3.2 ARM Algorithms

ARM was introduced by Agrawal et al. [8] to discover the occurrence of items in market transactions. The underlying concept is to verify $\mathcal{A} \rightarrow \mathcal{C}$ indicating if an item $\mathcal{A}$ exist as "antecedence", then item $\mathcal{C}$ should co-exist as "consequence". In order to identify the correlations of $\mathcal{A} \rightarrow \mathcal{C}$, let $\mathcal{D} = \{X_1, X_2, \ldots, X_n\}$ where $\mathcal{D}$ is the database, X_i is the ith instance, and n is the total number of instances from $\mathcal{N}$ in $\mathcal{D}$. For each X_i, there might be m number of items (i.e., features, attributes, factors) to generate frequent itemsets. The metrics for evaluating ARM algorithms are support and confidence values. The support value of each item in X_i can be calculated through Eq. 4.1, which is used to identify the frequency of an itemset.

$$Support = \frac{freq(\mathcal{A} \rightarrow \mathcal{C})}{\mathcal{N}} \tag{4.1}$$

When evaluating the frequent itemsets, the conditional probability $\mathcal{P}(\mathcal{C}|\mathcal{A})$ will also need to be paid attention through Eq. 4.2 to identify the confidence of an instance containing $\mathcal{A}$ and also $\mathcal{C}$.

$$Confidence = \frac{freq(\mathcal{A} \rightarrow \mathcal{C})}{freq(\mathcal{A})} \tag{4.2}$$

This thesis involved three ARM algorithms for identifying thyroid disease pathogenesis, including the two most classic ARM algorithms (i.e., Apriori and FP-Tree), as well as a proposed more advanced algorithm called Faster Apriori.

Apriori

The Apriori algorithm was initially proposed by Agrawal and Srikant [7]. The goal is to extract associations, frequent patterns, or even casual structures from unstructured datasets. The Apriori algorithm is relatively straightforward for implementation, which requires generating candidate itemsets for frequent rule sets identification. The detailed procedures are as follows:

- Pre-define thresholds for support and confidence values.
- Identify support values for all the individual items m from X_i, then prune the ones which do not meet the support threshold.
- Loop through iteration in $\mathcal{D}$, for each candidate m item in X_i, pare up until enumerating all items in X_i.
- Calculate support values for all the candidate itemsets and prune the ones below the threshold.
- Repeat the above two steps, each time including itemsets $m + 1$ for each X_i, until finishing listing all itemsets in $\mathcal{D}$.

- Final rules are the frequent itemsets with support and confidence values above thresholds.

Frequent Pattern-Growth Tree

The frequent pattern-growth (FP-Tree) was proposed by Han et al. [102], which is another classic algorithm that was adopted relatively often in the medical domain. Unlike the Apriori algorithm, the FP-Tree algorithm does not require the generation of candidate itemsets, making the rule extraction process more efficient when dealing with small-to-medium scaled datasets. The detailed procedures are as follows:

- Identify support values for all the individual items m in X_i.
- Write all the items m in descending order based on their support values.
- Draw the FP-tree starting from the "null" node and record the m items following the descending list.
- Update the FP-tree through each iteration; meanwhile, record and update the item frequency in the tree structure.
- Generate a conditional FP-tree if the support value for the node is larger than the pre-defined threshold.
- Generate frequent patterns based on the conditional FP-tree, and they are the final rules.

Faster Apriori

As the earliest ARM design, Apriori and FP-Tree are regarded as the two most classic algorithms to be implemented in miscellaneous disciplines today [258, 161, 159]. Although considered accurate in extracting valuable knowledge from multiple attributes, those two

algorithms are not computationally efficient enough when the dataset gets more extensive or complicated. Based on this, the Faster Apriori algorithm was proposed, which overcomes the efficiency drawback without harming the generate rules.

The Faster Apriori algorithm was developed to discover the occurrence of items in a large volume of digital records based on identifying probabilities for different itemsets initially and then updated to override the original transaction database to exclude redundant itemsets during each iteration. In order to extract valuable information, this research utilised one more metric to generate the final rules: the probability measure, calculated through Eqs. 4.3 and 4.4.

$$\mathcal{P}_{X_i} = \frac{freq(X_i^m)}{\mathcal{N}} \tag{4.3}$$

$$\mathcal{P} = \prod_{i=1}^{m} p(Y_i^m | X_i^m) \tag{4.4}$$

4.3.3 Exceptionality Measure

Exception rules are the ones that have low support values but high confidence values. The most well-known exception rule is "$Champagne \Rightarrow Caviar$", which does not have a high frequency in the database because they are pricey, but they are always brought together [296]. Exception rules can sometimes be influential and valuable. However, in the existing studies, researchers tend to put their efforts into identifying common rules, whereas exception rules were greatly neglected. This research incorporates the exceptionality measure to reveal exceptional underlying knowledge of thyroid disease pathogenesis.

Based on Piatetsky-Shapiro's arguments [80] and probability theory, the measurements of common rules and exception rules should be different. This thesis incorporates the conditional-probability increment ratio (CPIR) function proposed by Wu et al. [323] as an additional measurement for rules selection to evaluate the dependency of the antecedent A and the consequent $\mathcal{C}$. The CPIR score generates a number between -1 and 1. When

the CPIR score is positive, the items are positively related. On the other side, when the value is negative, the items are negatively related. More specifically, the CPIR function for common rules evaluation was carried out by Eq. 4.5 and for the exception rules through Eq. 4.6. With the adoption of CPIR for both types of rules, the dependency of the antecedent and the consequent can be evaluated and interpreted.

$$CPIR(X_i^m \rightarrow Y_j) = \frac{sup(X_i^m \bigcup Y_j) - sup(X_i^m) \times sup(Y_j)}{sup(X_i^m) \times (1 - sup(Y_j))} \tag{4.5}$$

$$CPIR(X_i^m \rightarrow \neg Y_j) = \frac{sup(X_i^m \bigcup \neg Y_j) - sup(X_i^m) \times sup(\neg Y_j)}{sup(X_i^m) \times sup(Y_j)} \tag{4.6}$$

4.3.4 Implementation Procedure

This thesis aims to extract common and exception rules simultaneously for thyroid disease pathogenesis identification. Therefore, with the thresholds defined for the CPIR measurements, the reliable common and exception rules can be identified. Moreover, the detailed procedure of the exception rules generation is shown in Algorithm 1.

Compared to the classic Apriori algorithm, the Faster Apriori algorithm proposed the probability-based frequent pattern mining approach to initially mitigate the computational resources for frequent itemsets generation [261]. Specifically, the initial $\mathcal{D}$ was utilised to identify unique items m from X_i by identifying the probability of the individual unique attribute. The probability of combinations for the unique $m+1$ during the iteration was also calculated and compared to the pre-defined threshold. Following the exclusion of under-qualified $m+1$ and the corresponding combination pairs, a new dataset $\bar{\mathcal{D}}$ was generated for frequent itemsets generation intuitively. Additionally, in order to reduce the computational cost for efficiency optimisation, the dataset $\bar{\mathcal{D}}$ was updated in each iteration when finding frequent itemsets to exclude redundant itemsets. Algorithm 2 demonstrates the proposed Faster Apriori implementation procedure.

Before generating candidate itemsets like the classic Apriori algorithm, the probability-

Algorithm 1: Pseudo-code for Exception Rules Generation

Input:
$\mathcal{D} = \{X_1, X_2, \ldots, X_n^m\}$; X is the patient instances and n is total number of cases
$\theta_{Sup} = (p, q]$, θ_{Con}; Set the threshold interval for support and set minimum confidence
Initialisation:
ARM algorithm: Apriori, FP-Growth, or Faster Apriori
Set minimum support $\theta_{Sup} = p$
$Rules \leftarrow (Sup, Con)$; Rule sets generated through ARM algorithm with minimum support
Set $i = 0$
while $i \leq len(Rules)$ **do**
  if $Con_i \geq \theta_{Con}$ **then**
    $Rules \leftarrow (Sup_i, Con_i)$; Rules greater than minimum confidence values will be stored
  else
    Remove Rules
  end
  Set $\theta_{Sup} = q$
  for i *in* $Rules$ **do**
    if $Sup_i \leq \theta_{Sup}$ **then**
      $Rules \leftarrow (Sup_i, Con_i)$; Rules in between support interval values will be stored
    else
      Remove Rules
    end
    Set θ_{CPIR}
    Calculate CPIR for each i in $Rules$ using Eq. 4.6
    for i *in* $Rules$ **do**
      if $CPIR_i \geq \theta_{CPIR}$ **then**
        $Rules \leftarrow (Sup_i, Con_i)$; Rules greater than threshold CPIR values will be stored
      else
        Remove Rules
      end
    end
  end
  $i = i + 1$
end
Store and plot all the final rules
Output: Final exception rules $\rightarrow Exp_{final}$

Algorithm 2: Faster Apriori With Exception Rules Extraction

Input:
$\mathcal{D} = X_{i\in(0,n)}^{m(\geq 1)}$; X is the instances, n is total number of cases, m is the feature vector
$\theta_{Prob}, \theta_{Sup}, \theta_{Con}, \theta_{CPIR}$; Set the threshold probability, support interval, minimum confidence, and minimum CPIR
Generate an initial itemset $item_{ini}$ with all unique items in X
Calculate $\mathcal{P}_X$ for all $item_{ini}$ with Eq. 4.3
Calculate $\mathcal{P}$ for paired combinations $item_{pair}$ of items in $item_{ini}$ with Eq. 4.4
Exclude $item_{ini}$ and $item_{pair}$ that are below θ_{Prob}
Produce a list of candidate itemsets C^m with remaining $item_{ini}$ and $item_{pair}$
Generate $\bar{\mathcal{D}}$ based on C^m
Initialisation:
Set $n = 0$
Set $itemset_{temp}$ to an empty list
while $n < \bar{N}$ **do**
  Set $i = 1$; i is the number of i items in $\bar{X}_n$
  while $i \leq len(\bar{X}_n^m)$ **do**
    if i *not in* C^m **then**
      i append to C^m
    end
    Calculate support values for all C^m using Eq. 4.1
    if $Sup_i > \theta_{Sup}$ **then**
      i append to $itemset_{freq}$;
      Update $\bar{\mathcal{D}}$ with $itemset_{freq}$ in place
    end
    $i = i + 1$
  end
  $n = n + 1$
end
Calculate confidence for all $itemset_{freq}$ use Eq. 4.2
Calculate CPIR for all $itemset_{freq}$ use Eq. 4.5
$Rules \leftarrow itemset_{freq}, Sup_i, Con_i, CPIR_i$
if $Rules \geq \theta_{Sup}, \theta_{Con}, \theta_{CPIR}$ **then**
  $Rules$ will be stored
end
Store and plot all the final rules
Output: $Final \leftarrow Rules(Sup, Con, CPIR)$; store all qualified rules with exceptions

based frequent patterns were generated using the probability measure $\mathcal{P}_X$ to identify the probability scores for all unique items in X. Then, the $\mathcal{P}$ score was measured to evaluate all the possible combinations with unique m. The generated unique m and its corresponding combinations $m+1$ were compared to the probability threshold. Then the under-qualified ones were removed, whereas the qualified ones were remained to generate a new $\bar{\mathcal{D}}$. This process eliminated the itemsets that were not frequently seen in the database, thus reducing the running time in the later iterations for frequent candidate itemsets generation. The support values were calculated for all the newly generated candidate itemsets. During each iteration, the database $\bar{\mathcal{D}}$ was updated to exclude those below the threshold. In this way, the old database was overridden to reduce computational needs in each iteration. The meaningless itemsets would not be considered for rules generation, thus enhancing the efficiency to an optimised level. Then, the evaluations of the frequent itemsets after finishing all the iterations were the confidence and CPIR values to further select the most valuable factors.

4.4 Experimental Design

This chapter separately evaluates the risk factors of thyroid disorder (i.e., hypothyroidism and hyperthyroidism) and thyroid cancer. This section illustrates the adopted datasets and the experimental settings.

4.4.1 Dataset Descriptions

This section includes two datasets to evaluate the proposed TM-DM framework for knowledge extraction. The two datasets are described as follows.

Open-access UCI Dataset

The open-access dataset was adopted in this chapter to identify risk factors for thyroid disorders, which was retrieved from the UC Irvine (UCI) Machine Learning Repository

[73]. This dataset has been pre-processed into *.csv* format, containing 21 attributes and 7,200 instances with thyroid disorder-related diagnoses. After a set of data wrangling processes (i.e., data cleaning, duplicates removal, data transformation), a total number of 5,600 instances were utilised in this section. The selected attributes can be found in Table 4.1.

Table 4.1: UCI thyroid disorder dataset attributes.

Attributes	**Descriptions**
Age	Age group intervals
Sex	M=Male and F=female
I131	Had I131; f=False and t=True
Sick	Sick status; f=False and t=True
Psych	Have psych; f=False and t=True
Goitre	Have goitre; f=False and t=True
Tumour	Have tumour; f=False and t=True
Lithium	Lithium status; f=False and t=True
Thyroxine	Take thyroxine; f=False and t=True
Pregnant	Pregnant status; f=False and t=True
Query hypothyroid	Hypothyroidism; f=False and t=True
Query hyperthyroid	Hyperthyroidism; f=False and t=True
Thyroid surgery	Had thyroid surgery; f=False and t=True
Hypopituitary	Hypopituitary statue; f=False and t=True
Anti-thyroid medication	Antithyroid medication; f=False and t=True
TSH	TSH level; numerical value
$T3$	$T3$ level; numerical value
$TT4$	$TT4$ level; numerical value
$T4U$	$T4U$ level; numerical value
FTI	FTI level; numerical value
Class	Thyroid disorder; negative and positive

Self-acquired CN Hospital Dataset

This chapter also involves the self-acquired dataset for determining the risk factors associated with thyroid cancer. We obtained 578 in-patient digital health records from a first-class Chinese hospital (CN Dataset) with ethics approval from Monash University.

Those patients have been diagnosed with thyroid-related diseases (e.g., goitre, adenoma, cancer) from August 2018 to August 2021. The obtained records include patients' admission reports, diagnostic reports, and discharge summaries. Those reports

were stored in *.pdf* format in the electronic health record (EHR) system. Therefore, text mining procedures were incorporated to extract key terminologies from those medical records. The extracted terms were then adopted and transformed into a *.csv* file, indicating whether the risk factor was present in the patient records.

More specifically, the raw admission reports and discharge summaries were used for critical terminologies extraction. The admission reports contain the patient's demographic information, medical history, lifestyle behaviours, and current symptoms. The discharge summaries include the patient's treatment protocols, comorbidities, and principal diagnosis. The extracted attributes were then normalised through stemming and lemmatisation following the text mining procedures. In order to reach consistency, the extracted information included not only the same attributes from the UCI repository but also the history of diseases, comorbidity, and principal diagnosis. Moreover, diagnostic reports were also used to extract patients' blood examination levels. Therefore, 32 attributes were adopted for the hospital dataset, and details are available in Table 4.2.

Table 4.2: Self-acquired CN hospital thyroid cancer dataset attributes.

Attributes	**Descriptions**
Obesity	Have obesity; f=False and t=True
Diabetes	Have diabetes; f=False and t=True
Radiation	Had radiation; f=False and t=True
Depression	Have depression; f=False and t=True
Hypertension	Have hypertension; f=False and t=True
Vitamin D deficiency	Have VD deficiency; f=False and t=True
$FT3$	$FT3$ level; numerical value
$FT4$	$FT4$ level; numerical value
$TGII$	$TGII$ level; numerical value
$TGAb$	$TGAb$ level; numerical value
$TPOAb$	$TPOAb$ level; numerical value
Class	Thyroid cancer; negative and positive

4.4.2 Experimental Settings

For the UCI dataset, the final list of attributes and instances were selected based on the following mechanism:

- All the instances with missing age were removed.
- All the instances with missing gender were removed.
- All the categorical variables were transformed into numerical values.
- Numerical variables with missing values were assigned random numbers between normal ranges of the blood examinations: TSH: 0.27 - 4.2, $T3$: 1.3 - 3.1, $TT4$: 62 - 164, $T4U$: 0.7 - 1.8, FTI: 53 - 142.

The minimum support threshold for the ARM algorithms implementation was set to 0.5. The minimum confidence threshold was set as 0.7 for common rules extraction. As far as the exception rules generation was concerned, the support interval was set to $(0.2, 0.4]$ with the same minimum confidence as common rules.

On the other hand, for thyroid cancer pathogenesis identification, the self-acquired CN dataset selects attributes through the mechanism as follows:

- Principal diagnosis missing or unclear were removed.
- Risk factors present denoted with 1, otherwise absent as 0.
- Numerical variables with missing values were assigned random numbers between normal ranges of the blood examinations: $FT3$: 3.6 - 7.5, $FT4$: 12 - 22, $TGII$: 3.5 - 77, $TGAb$: 11 - 115, $TPOAb$: 0 - 34.

For the self-acquired dataset, the minimum support value was set a bit lower than the UCI data settings because the scale of the dataset was relatively small. Therefore, the minimum support value and confidence threshold for common rules were 0.6 and 0.7, respectively. As for the exception rules, the support interval was also set to $(0.2, 0.4]$, and the confidence threshold was still 0.7. Besides, the minimum CPIR score for the UCI dataset was set to 0.1, the hospital dataset was set to 0.2, and the final rules were sorted by confidence and CPIR values.

4.5 Results

The proposed TM-DM framework was evaluated following the correlation analysis, feature selection, and classification procedures. The final classification performance evaluates the pre-defined hypotheses for thyroid disease pathogenesis identification.

4.5.1 ARM Selected Attributes

The mutual rules extracted from the three ARM algorithms were described and recorded in this section based on the different gender and disease groups. Besides, the conflicting rules for the different algorithms were removed following forward and backward reasoning. In addition, the retained rules are reliable ones which can be found in all the algorithms. The extracted final rules are presented as follows.

Thyroid Disorder Risk Factors

Table 4.3 presents the extracted top three association rules for thyroid disorder patients from different gender groups using the UCI dataset. For the extracted common and exception rules generated through the ARM algorithms, it is evident that age, TSH, $T3$, and FTI levels are directly correlated with various types of thyroid disorders.

Regardless of gender, patients aged from 40 to 80 were more likely to be diagnosed with thyroid disorders. Based on the common rules for the male group, patients aged from 40 to 60 with abnormal thyroid-stimulating hormone levels are likely to be diagnosed with hypothyroidism. This rule was also found in the female group. Moreover, females aged from 60 to 80 are also likely to establish thyroid disorders. Similar patterns were found in the generated exception rules.

It should be paid attention to the female groups from the exception rules aspect. Besides TSH and $T3$, the FTI level became a critical indicator, with a CPIR score of 0.73, indicating the association between this blood index measurement and thyroid disorder is relatively strong.

Table 4.3: Open-access UCI thyroid disorder dataset - extracted rules.

Common Rules					
Groups	**Association Rules**		**Class**	**Conf.**	**CPIR**
Male	$Age = (40, 60]$, $TSH = Abnormal$	$\rightarrow$	$Disorder$	1.00	1.00
	$TSH = Abnormal$, $T3 = Abnormal$	$\rightarrow$	$Disorder$	1.00	1.00
	$TSH = Abnormal$	$\rightarrow$	$Disorder$	0.88	0.11
Female	$Age = (60, 80]$	$\rightarrow$	$Disorder$	1.00	1.00
	$Age = (40, 60]$, $TSH = Abnormal$	$\rightarrow$	$Disorder$	0.99	0.70
	$TSH = Abnormal$	$\rightarrow$	$Disorder$	0.75	0.12
Exception Rules					
Male	$Age = (60, 80]$, $TSH = Abnormal$	$\rightarrow$	$Disorder$	0.91	0.34
	$Age = (60, 80]$	$\rightarrow$	$Disorder$	0.91	0.34
	$Age = (40, 60]$	$\rightarrow$	$Disorder$	0.88	0.12
Female	$FTI = Abnormal$	$\rightarrow$	$Disorder$	0.83	0.73
	$Age = (60, 80]$, $TSH = Abnormal$	$\rightarrow$	$Disorder$	0.76	0.16
	$TSH = Abnormal$, $FTI = Abnormal$	$\rightarrow$	$Disorder$	0.75	0.61

Thyroid Cancer Risk Factors

Table 4.4 demonstrates the extracted association rules for thyroid cancer patients using the self-acquired CN dataset. Based on the table, thyroid cancer-related factors are age, gender, TSH, $T3$, $FT3$, $FT4$, $TGAb$, thyroxine intake status, history of thyroid disease, thyroid surgery history, and tumour history.

For the extracted common rules, patients with abnormal thyroid hormone levels should be examined for thyroid cancer development despite the age factor. In particular, male patients with tumour history and female patients with thyroid-related surgery in the past should pay close attention to the disease.

As far as the generated exception rules are concerned, male patients aged between 18 to 25 and 40 to 60 should check thyroid hormones regularly. Female patients with hypothyroidism in the past are likely to establish subsequent thyroid cancer.

4.5.2 FS Selected Attributes

In order to evaluate the extracted risk factors, a comparative analysis was performed through a set of feature selection algorithms. Four classic statistical feature selection

Table 4.4: CN thyroid cancer dataset - extracted rules.

Common Rules					
Groups	**Association Rules**		**Class**	**Conf.**	**CPIR**
Male	$FT3, FT4 = Abnormal$	$\rightarrow$	$Cancer$	0.75	1.00
	$TSH, FT3, FT4 = Abnormal$	$\rightarrow$	$Cancer$	0.75	1.00
	$FT3, FT4 = Abnormal, Tumour$	$\rightarrow$	$Cancer$	0.75	1.00
Female	$FT4 = Abnormal, Thyroxine$	$\rightarrow$	$Cancer$	1.00	1.00
	$FT4 = Abnormal, Thy\_surgery$	$\rightarrow$	$Cancer$	1.00	1.00
	$TSH = Abnormal, Thyroxine, Thy\_surgery$	$\rightarrow$	$Cancer$	0.93	0.63
Exception Rules					
Male	$Age = (18, 25], TSH, FT3, FT4 = Abnormal$	$\rightarrow$	$Cancer$	1.00	1.00
	$Age = (18, 25], FT4 = Abnormal, Tumour$	$\rightarrow$	$Cancer$	1.00	1.00
	$Age = (40, 60], T3, TGAb = Abnormal$	$\rightarrow$	$Cancer$	1.00	1.00
Female	$TGAb = Abnormal$	$\rightarrow$	$Cancer$	1.00	1.00
	$FT4 = Abnormal, Hypothyroidism$	$\rightarrow$	$Cancer$	1.00	1.00
	$Age = (25, 40], TSH, FT3 = Abnormal$	$\rightarrow$	$Cancer$	1.00	1.00

algorithms were involved in this case. Table 4.5 demonstrates the selected features using Chi, MI, fisher, KW, and ARM algorithms.

For the UCI thyroid disorder dataset, the common attributes selected from the feature selection algorithms were age, gender, thyroxine, TSH, $T3$, and FTI. For the CN thyroid cancer dataset, the common attributes among the five algorithms were age, gender, thyroxine, thyroid surgery history, hypertension, TSH, $T3$, $FT3$, and $FT4$ levels. Those selected attributes were then validated through a classification task, and the results are presented in the next section.

4.5.3 Classification Performance

Table 4.6 demonstrates the classification performance of the baseline (i.e., with all features) and feature selection algorithms in the six classifiers. Through the 10-fold CV, the standard deviation scores of each algorithm in each classifier were also presented.

For the UCI dataset, the best-performing model for the baseline was the RF classifier with an accuracy of 0.903. The competitive performance was achieved through the MI and the Fisher score algorithms with the same classifier. Moreover, the ARM algorithms selected features reached the same accuracy by employing a 3 hidden neurons MLP model.

Table 4.5: FS and ARM Selected features.

UCI Dataset		
Methods	**No.**	**Selected Features**
Baseline	20	-
Chi	10	Sex, Thyroxine, Pregnant, Hypothyroid, Hyperthyroid, Tumour, TSH, $T3$, $TT4$, FTI
MI	16	Age, Sex, Thyroxine, Sick, Pregnant, Thyroid_surgery, Hypothyroid, Goitre, Tumour, Hypopituitary, Psych, TSH, $T3$, $TT4$, $T4U$, FTI
Fisher	18	Age, Thyroxine, Antithyroid_medication, Sick, Pregnant, Thyroid_surgery, $I131$, Hypothyroid, Hyperthyroid, Lithium, Goitre, Tumour, Hypopituitary, Psych, TSH, $T3$, $TT4$, $T4U$
KW	10	Age, Sick, Hypothyroid, Hyperthyroid, Psych, TSH, $T3$, $TT4$, $T4U$, FTI
ARM	6	Age, Sex, Thyroxine, TSH, $T3$, FTI
CN Dataset		
Baseline	30	-
Chi	10	*Age*, Thyroxine, Thyroid_surgery, $I131$, Tumour, $T3$, $TT4$, $FT3$, $FT4$, $TGII$
MI	15	Age, Thyroxine, Sick, Thyroid_surgery, $I131$, Tumour, Psych, TSH, $T3$, $TT4$, $FT3$, $FT4$, $TGII$, Hypertension, Diabetes
Fisher	28	Age, Sex, Thyroxine, Antithyroid_medication, Sick, Pregnant, Thyroid_surgery, $I131$, Hypothyroid, Hyperthyroid, Lithium, Goitre, Tumour, Hypopituitary, Psych, TSH, $T3$, $TT4$, $FT3$, $FT4$, $TGII$, $TGAb$, $TPOAb$, Vitamin_D_Deficiency, Hypertension, Diabetes, Depression, Obesity
KW	11	Antithyroid_medication, Sick, $I131$, Hyperthyroid, Goitre, Psych, $TT4$, $TPOAb$, Vitamin_D_Deficiency, *Depression*, Obesity
ARM	11	Age, Sex, Thyroxine, Thyroid_surgery, Hypothyroid, Tumour, TSH, $T3$, $FT3$, $FT4$, $TGAb$

The ARM models all generated similar classification performance compared to other feature selection models with LR, DT, and NB classifiers. However, a minor decrease was found when applying SVM and RF classifiers compared to the baseline.

With respect to the CN dataset, the best-performing model for the baseline was LR, with an accuracy of 0.962. A competitive performance was also achieved with Chi, MI, and KW algorithms. The ARM-based features tend to perform well through DT, SVM, RF, and NB classifiers compared to the baseline. The best-performing model is LR with the Kruskal-Wallis feature selection algorithm, which obtained an averaged accuracy of 0.965 for the 10-fold CV. Based on the analysis, the ARM techniques tend to select the least number of attributes while obtaining competitive classification accuracy rates.

Table 4.6: Classification performance with feature selection (10-fold CV).

	UCI Dataset					
	LR	**DT**	**SVM**	**RF**	**NB**	**MLP**
Baseline	$0.898 \pm .01$	$0.810 \pm .06$	$0.897 \pm .02$	$0.903 \pm .01$	$0.897 \pm .03$	$0.898 \pm .02$
Chi	$0.899 \pm .01$	$0.810 \pm .06$	$0.896 \pm .02$	$0.900 \pm .01$	$0.897 \pm .03$	$0.903 \pm .01$
MI	$0.899 \pm .01$	$0.810 \pm .06$	$0.896 \pm .02$	$0.903 \pm .01$	$0.897 \pm .03$	$0.900 \pm .01$
Fisher	$0.900 \pm .01$	$0.810 \pm .06$	$0.901 \pm .01$	$0.903 \pm .01$	$0.897 \pm .03$	$0.894 \pm .01$
KW	$0.900 \pm .01$	$0.808 \pm .06$	$0.897 \pm .02$	$0.901 \pm .01$	$0.896 \pm .03$	$0.902 \pm .01$
ARM	$0.900 \pm .01$	$0.810 \pm .06$	$0.895 \pm .02$	$0.898 \pm .02$	$0.897 \pm .03$	$0.903 \pm .01$
	CN Dataset					
Baseline	$0.962 \pm .03$	$0.930 \pm .07$	$0.935 \pm .01$	$0.957 \pm .03$	$0.892 \pm .08$	$0.957 \pm .04$
Chi	$0.957 \pm .04$	$0.941 \pm .07$	$0.943 \pm .03$	$0.962 \pm .03$	$0.849 \pm .09$	$0.941 \pm .03$
MI	$0.957 \pm .03$	$0.941 \pm .06$	$0.962 \pm .03$	$0.960 \pm .03$	$0.919 \pm .04$	$0.962 \pm .03$
Fisher	$0.960 \pm .04$	$0.941 \pm .07$	$0.935 \pm .01$	$0.960 \pm .03$	$0.900 \pm .08$	$0.946 \pm .04$
KW	$0.965 \pm .03$	$0.954 \pm .04$	$0.935 \pm .01$	$0.962 \pm .03$	$0.768 \pm .17$	$0.952 \pm .02$
ARM	$0.960 \pm .03$	$0.949 \pm .05$	$0.935 \pm .01$	$0.960 \pm .02$	$0.903 \pm .07$	$0.949 \pm .03$

4.6 Discussion

Although thyroid disease is prevalent in the contemporary, the underlying cause of the disease remains unclear. ARM technique has been applied relatively often in the medical domain, whereas the implementation to identify the pathogenesis of thyroid disease is absent. Besides, generating rules directly from raw digital health records and identifying

exception rules are the novelties in this study.

The proposed TM-DM framework was analysed by adopting two sources of thyroid disease-related records, where the results confirmed that sex and age are the two leading factors correlated to thyroid disease. This finding aligns with the existing works in the literature [253, 110], and it might be due to hormonal factors, including the impact of pregnancy or pubertal development; these are particularly sensitive to young females.

Besides age and gender, the generated results from both datasets manifest that a history of thyroid-related diseases, like hypothyroidism, hyperthyroidism, or past thyroid surgery, increases the risk of establishing subsequent thyroid cancer. This finding is intriguing, and matches with the existing study [64]. The finding confirmed that a subsequent thyroid cancer risk was highly enhanced if thyroid disease existed in the past. Therefore, the history of thyroid diseases can be a good indicator when diagnosing current thyroid status.

Comorbidities like diabetes, obesity, hypertension, depression, psychiatric diseases, and vitamin D deficiency were also included for evaluation. Among all the factors, psychiatric diseases are not strongly related to the pathogenesis of thyroid disease. For the other factors, the results exhibit no solid positive associations were found between the comorbidities with thyroid disease. This result is in accordance with Shih et al. [277], but controversial with Ma et al. [185]. Nevertheless, it was found that the absence of those underlying health problems like obesity, depression, hypertension, and diabetes will reduce the risk of being diagnosed with thyroid cancer. Besides, vitamin D deficiency might be influential to thyroid disease, and this finding aligns with Zhao et al. [346]. However, further evaluations should be involved to ascertain the associations since the sample scale is relatively limited in this study.

4.7 Summary

The proposed TM-DM framework was used to identify and evaluate the risk factors correlated with thyroid disease, including thyroid disorder and thyroid cancer. Through the utilisation of two data sources, the common and exception association rules were extracted independently. For thyroid disorder, the leading factors were found to be age, gender, thyroxine intake status, and thyroid function examinations, like TSH, $T3$, and FTI. For thyroid cancer, the leading factors were identified, including the history of thyroid disease, hypertension, the history of thyroid surgery, $FT3$, and $FT4$ levels besides age, gender, thyroxine, TSH, and $T3$ levels.

Nevertheless, this research study did not confirm positive associations between controversial factors like vitamin D deficiency, diabetes, and depression with thyroid cancer. Additionally, during the classification evaluation stage, the performance among the different feature selection algorithms was comparable; this might be due to the limited number of features incorporated. Therefore, the alternative research direction would enhance the scale of the sample, include more attributes for evaluations, and include as many controversial factors as possible, such as gene heredity, mutations, and hormonal factors, to determine their associations with thyroid cancer.

Hence, this study emphasises the contributions made to society. Identifying the correlated risk factors allows thyroid disease mortality and morbidity rates to be mitigated considerably. In addition, the proposed TM-DM framework can be generalised to different diseases for more precious knowledge discovery in the medical domain, more importantly, strengthening the use of precision medicine or treatments to avoid certain diseases.

Chapter 5

Deep Convolutional Neural Networks in Thyroid Disease Detection

5.1 Introduction

The clinical diagnostic procedure for thyroid disease is relatively fussy and inefficient. Patients usually undergo a set of examinations such as thyroid function tests, medical image scanning (e.g., ultrasound, CT, MRI, radio-iodine scintigraphy, or positron emission tomography), FANC, or even biopsy to arrive at an accurate diagnostic decision. However, all these examinations are correlated with varying degrees of uncertainties in human false-positive and false-negative rates. Therefore, streamlining the process has the potential to increase the accuracy and efficiency of diagnostic decision-making.

In recent years, deep learning techniques, specifically convolutional neural networks (CNN), have improved diagnostic performance in interpreting medical images. Specifically, existing studies have put much effort into engaging with ultrasound images to detect thyroid cancer [50, 327, 146, 221], all demonstrating superior performance compared to radiologists. Ultrasonography is non-invasive and cost-efficient, and it can provide detailed structures of thyroid nodules, making it well-accepted for thyroid suggestive of malignancy [201, 243, 289]. Nevertheless, the process is operator-dependent, which may

result in inter-observer variations, much less say that it is highly susceptible to noises and speckles [309]. Moreover, the existing CAD models are usually designed to classify thyroid nodules individually, which is highly inefficient, leading to absent implementations in the clinical domain.

Besides ultrasonography, the CT scan is also recommended when diagnosing thyroid disease, whereas it is always suggested prior to surgeries for evaluating central lymphatic metastasis [139]. CT scan is comparatively more consistent than ultrasonography as it has fewer human factors influencing the image quality. Additionally, CT is influential in defining locations of abnormal thyroid nodules, relationships among structures, malignant invasion, and extent of retrosternal extension [351]. Relying on any of the unitary image modalities for diagnostic decision-making is not convincing enough. Therefore, the two imaging modalities are complementary, and their comparison of CNN performance for thyroid disease detection was made in this chapter.

Based on the provided comprehensive literature review analysis, most existing CAD models were designed to distinguish between benign and malignant thyroid nodules regardless of the adopted image modalities. In practice, functional and neoplastic thyroid diseases undergo varied treatment protocols. However, the existing CAD models, thought to be efficient, still have limitations in generating expert-level diagnoses. Accordingly, sub-classifying thyroid disease types should be achieved for precise treatments, potentially enhancing clinical applications of deep learning algorithms.

Therefore, in this chapter, the use of CAD models was described to mitigate human false-positive and false-negative diagnostic rates through binary classification tasks and achieve precise diagnosis through multi-classification tasks, using ultrasound and CT images for comparative analysis. The CAD models can assist clinicians in streamlining patient management and diagnostic decision-making. It can also help to avoid unnecessary FNAC cytology or biopsies of non-suspicious lesions, potentially mitigating patients' physical and financial stress. The related background words of this chapter were described in Chapter 2.3.2.

The contents of this chapter have been published in journal articles 1, 2.

5.2 Problem Formulation

In order to reach precise diagnoses for thyroid disease, the binary and multi-class classification tasks were implemented in this chapter with the use of ultrasound and CT images for comparative evaluation.

Given the image dataset $\mathcal{D} \in \{X, y\}$, let $X = \{X_1, X_2, \ldots, X_n\}$ where X_i is the ith image matrix and n is the total number of images. Specifically, $X_i \in \mathcal{R}_i^{w \times h \times c}$, where the images are denoted as w width and h height in RGB channels. Additionally, $y \in \{0, 1\}$ is used for the binary classification tasks, where 0 indicates the image is labelled as normal gland and 1 as abnormal. For the multi-classification tasks, $y \in \{0, 1, 2, 3, 4, 5\}$ where the image is labelled as 0 - normal, 1 - thyroiditis, 2 - cystic nodular, 3 - multi-nodular goitre, 4 - adenoma, and 5 - cancer, respectively. Those six classes were pre-defined and selected based on their treatment protocols. With the objective to reach consistency for CNN evaluation, the ultrasound and CT images were all resized into $X_i \in \mathcal{R}_i^{224 \times 224 \times 3}$.

With the goal to enhance thyroid disease diagnostic performance, the classic and advanced CNN models were evaluated in this chapter, including VGG [280], ResNet [105], Inception [293], DenseNet [114], and Xception [57]. It should also be noted that all these models were used for the binary classification task, and the top three best-performing models were applied for the multi-classification task for efficiency.

5.3 Methodology

In order to obtain automatic diagnoses for thyroid disease through deep learning techniques, a rigorous procedure has been applied for CAD designs incorporating two image modalities. This section describes the overall implementation process and the selected CNN models.

The Proposed CNN-BM Framework

Figure 5.1 depicts the CNN implementation process for detecting thyroid disease in binary and multi-class classification tasks. More specifically, the ultrasound and CT image sets were pre-processed to ensure consistency. In particular, all the images were re-sized into 224×224 in resolution. The CT scans were segmented into left and right sides in the middle of the trachea through the generated python-based CT-segmentation tool (available in Appendix B). The reason behind segmenting CT scans is that patients might have two sides of the thyroid gland diagnosed with different diseases. Thus, segmenting the CT scan allows applying a different diagnostic label to each lobe individually.

Then, the labelling process was conducted manually and rigorously. All images were labelled based on the TIRADS score, the cytological examinations, and histopathological results. In practice, at least two pathologists are generally involved in the diagnostic decision-making process for each patient, and additional pathologists might also get involved in generating final decisions if disagreements occur between the two. There were chances that multiple diagnoses appeared in one image. Therefore, a dominant class was assigned to that corresponding image. The dominant class was defined based on the severity of the disease, following the sequence of the normal kind being the least severe kind, then thyroiditis, cystic nodule, multi-nodular goitre, adenoma, and cancer being the most severe type. Images without any TIRADS, cytology, or histopathology confirmations were excluded from this study.

The selected CNN models were trained and evaluated with the labelled images through k-fold cross-validation. In order to avoid over-fitting, the fine-tuning process was also applied during model training and validation. In this case, the top three best-performing CNN architectures were adopted for multi-classification tasks. The generated performance from ultrasound and CT scans were compared based on a series of evaluation metrics, including accuracy, precision, recall, F1, and NPV. Lastly, the model with the best performance was recommended for CAD implementations in the clinical domain, and the comparison between the two image modalities was also interpreted.

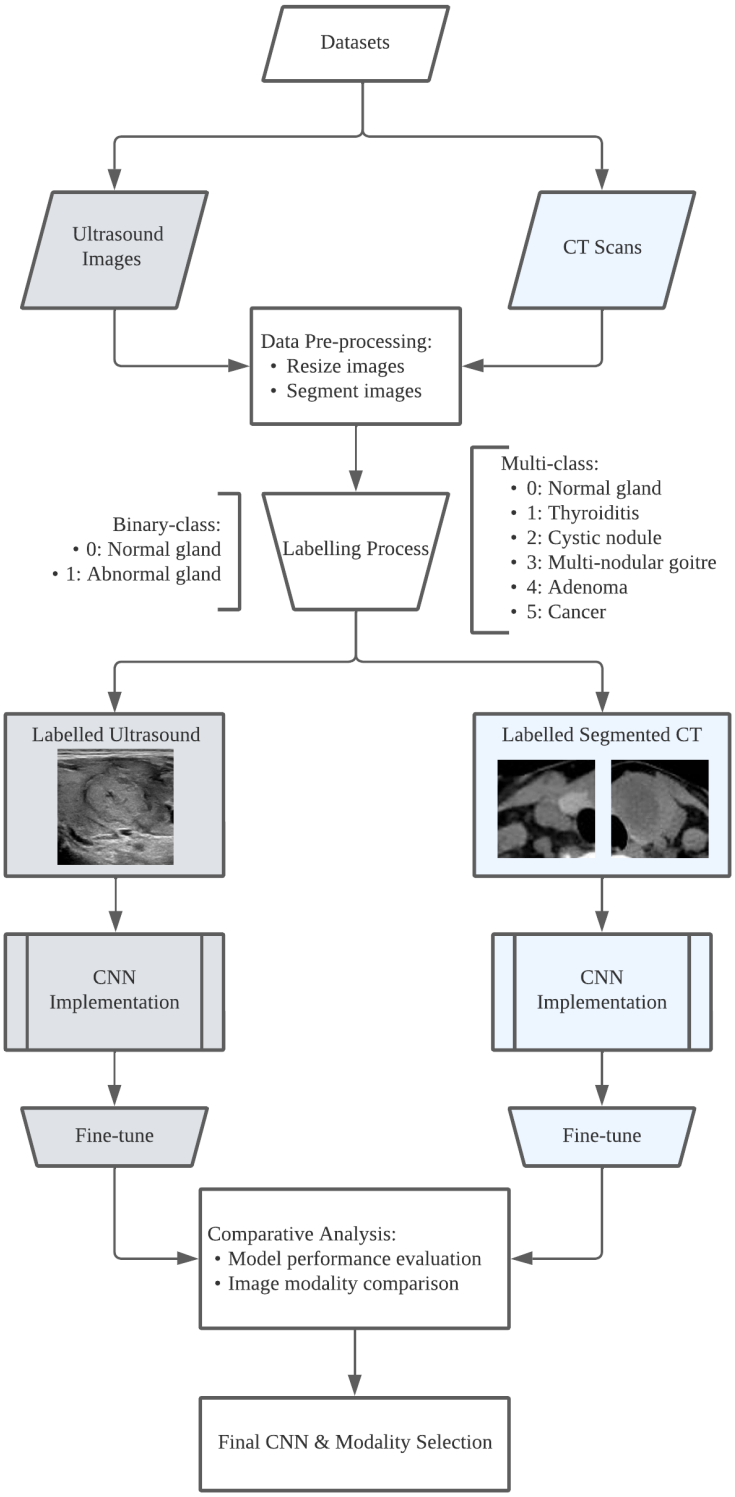


Figure 5.1: CNN-based binary & multi-class classification framework (CNN-BM) for thyroid disease diagnosis.

5.3.1 Network Architectures

With the emergence of deep learning techniques, the use of CNN models in the thyroid domain is relatively abundant, especially the use of VGG [72, 78, 260, 140] and ResNet models [97, 164, 70]. Those CAD models have demonstrated satisfactory diagnostic accuracy for thyroid disease detection ranging from 70% to 92%.

In order to have a comprehensive investigation of the CNN models in thyroid disease diagnosis, more advanced architectures were involved in this study, including Inception, DenseNet, and Xception models. More specifically, 11 CNN models were adopted and evaluated with the binary classification tasks, including VGG8, VGG11, VGG16, VGG19, ResNet10, ResNet18, ResNet50, DenseNet121, InceptionV3, InceptionResNetV2, and Xception.

With the development of the first CNN model, Le-Net, introduced in 1998 by LeCun et al. [156], AlexNet [149] and VGG [280] models were developed for classification purposes. Those models have standardised architectures that stack several convolution operations, following max-pooling or average-pooling layers and ending with fully connected layers. Those CNNs select critical features through the receptive field and are activated using the rectified linear unit (ReLU) function (Eq. 5.1) [275, 191]. The feature map size is generated through Eq 5.2. Here, $F(S)$ is used to represent the size of the feature map, n^w, n^h, n^c denote the input feature map size with width, height, and channel numbers, s stands for the stride number, while f represents the kernel size of the convolutional operation.

$$ReLU(x) = max(0, x) = max(0, \sum_{i=1}^{i=n} w_i x_i + b) \tag{5.1}$$

$$F(S) = (\frac{n_w - f}{s} + 1) \times (\frac{n_h - f}{s} + 1) \times n_c \tag{5.2}$$

Such architectures are generic and can be adjusted based on specific tasks, and the

commonly used architectures for VGG are VGG11, VGG16, and VGG19. However, researchers raised a concern that those models might not be appropriate to interpret medical images as they are much simpler in structures compared to natural images [34]. Accordingly, we reduced the depth of VGG architectures and proposed VGG8.

In theory, the more convolutional operations were stacked to the CNN, the better the classification results of the model would be. Nevertheless, the reality is that when more convolutions are stacked, the model will likely get a gradient explosion that can no longer implement the required tasks. In this regard, ResNet was proposed to address the issue as it can generate very deep CNNs, avoiding aggregating parameters that take exponentially increased computational resources [105]. More importantly, ResNet produces better classification results than conventional convolutional stacking CNN architectures as it can learn residuals through layers. This research proposes the ResNet10 architectures to compare with ResNet18 and ResNet50.

DenseNet was proposed by Huang et al. [114] to increase the accuracy caused by the vanishing gradient. The concept behind the model is that the information vanishes gradually before reaching the destination. Accordingly, the DenseNet was proposed to address the issue, which contains several dense blocks. Each layer from the dense block connects to all its preceding layers as the input so that the accuracy can be enhanced by reducing information loss.

The Inception model was proposed by Szegedy et al. [293], which introduced the inception module (Figure 5.2) to generate more accurate results. The inception module concatenates feature maps generated through different kernel size convolutions. Kernel size is influential to CNN performance, and classic CNN models apply fixed kernel size. In contrast, the inception module applies different kernel sizes simultaneously, including 1×1 to mitigate the feature map depth, 3×3 and 5×5 to obtain different information learned through the varied size of the receptive fields. The feature maps generated from the three kernels would be concatenated and fed into the next module, allowing enhanced accuracy as more information can be maintained during each convolutional operation.

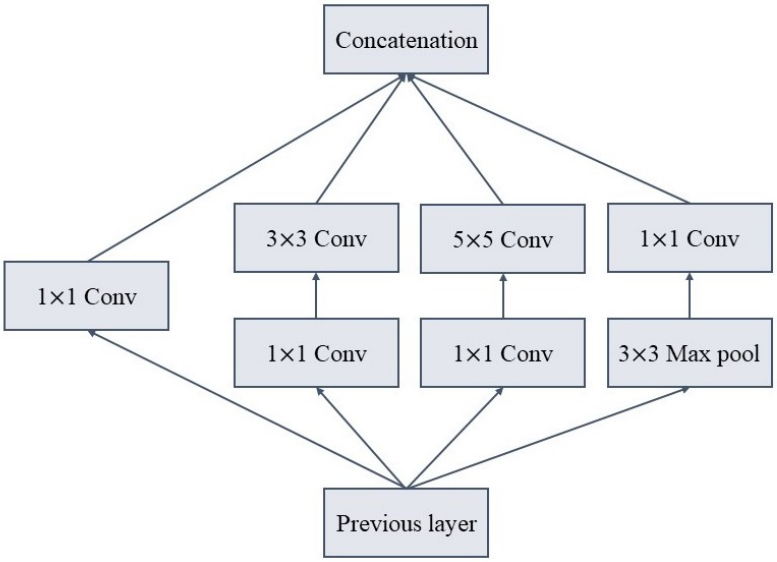


Figure 5.2: Inception module (adapted from [293]).

Xception was inspired by the Inception model [57], which was designed as the "extreme" version of the Inception module (Figure 5.3). Xception (Figure 5.4) maps cross-channel correlations from the input image and addresses spatial correlations of each output channel separately through the depth-wise separable convolutions. Chollet [57] once adopted the ImageNet database to evaluate Xception, and it reached the best accuracy and efficiency rates compared to VGG16, ResNet152, and InceptionV3 models.

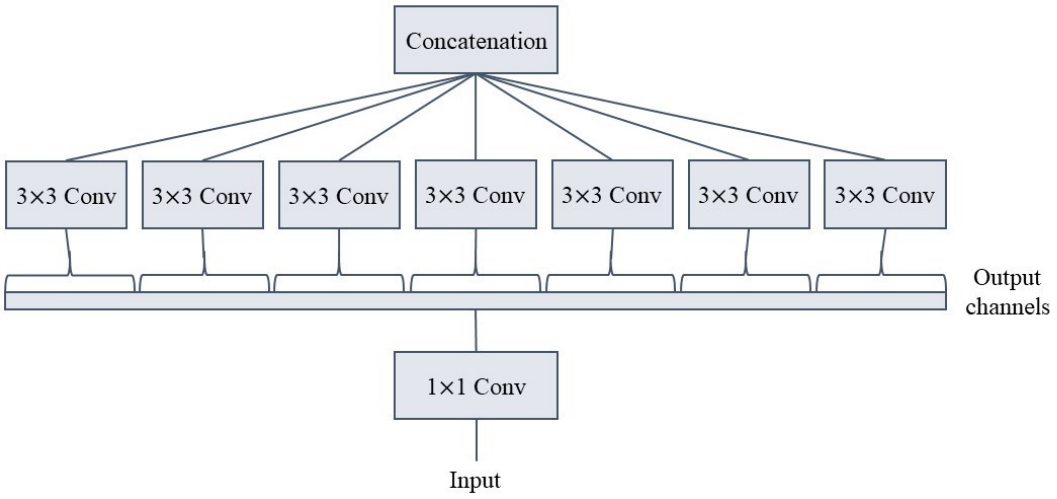


Figure 5.3: Extreme version of Inception module (adapted from [57]).

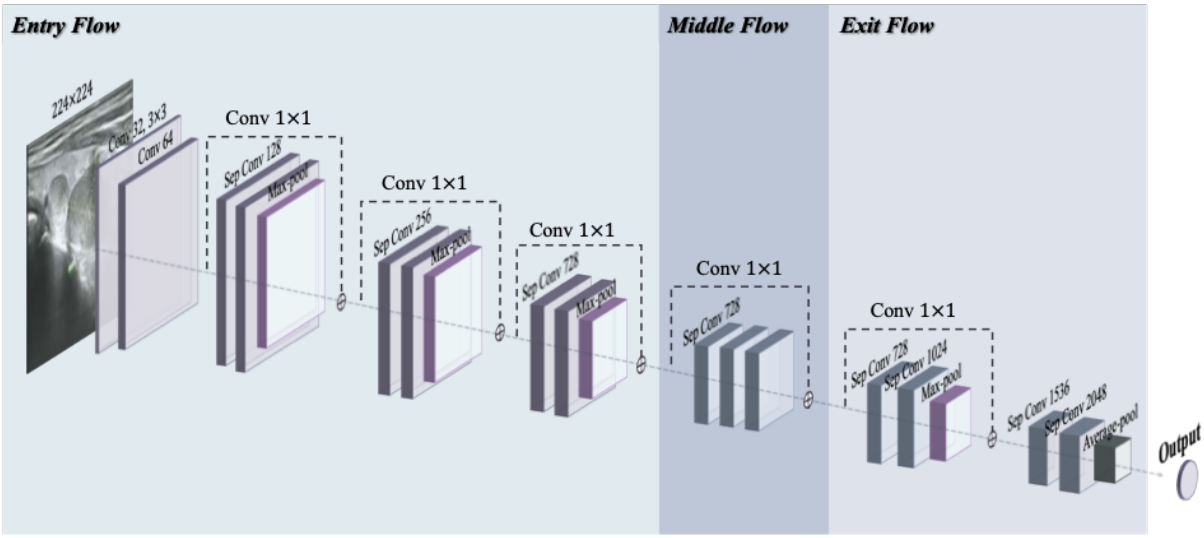


Figure 5.4: Xception architecture (adapted from [57]).

5.4 Experiments

This chapter evaluates the commonly used CNN architectures for thyroid disease detection through a binary classification task. The best-performing three models are further selected for the multi-class classification task. This section describes the utilised datasets and the parameter settings during the implementation.

5.4.1 Datasets Descriptions

With the ethical approval obtained, consecutive patients treated for thyroid diseases were recruited from a first-class hospital in Sichuan Province, China (CN dataset) between August 2018 and August 2021. The acquired data included de-identified radiological images and diagnostic reports from 578 patients. The distribution of demographic information of the dataset is presented in Table 5.1. Most patients were aged between 55 and 75 with a percentage of 42.54%. A substantial portion of the patients were females, with a percentage of 76.22%. Over 80% of the pathological results turned out to be benign, including thyroiditis, cystic nodule, multi-nodular goitre, and adenoma.

Table 5.1: Distribution of demographics of CN dataset.

Demographics	Percentage (%)
Age	
Below 18	0.35
18 - 35	10.24
35 – 55	40.45
55 – 75	42.54
75 +	6.42
Gender	
Male	23.78
Female	76.22
Pathology	
Benign	81.78
Malignant	18.22

Ultrasound Datasets

The CN dataset is acquired from the earliest developed hospital on the northern side of Sichuan province, with over 30 departments. The thyroid department was founded several years ago and has cured more than $10,000$ patients.

In practice, a complete assessment of the thyroid gland by ultrasonography generates a set of images demonstrating one nodule from diversified angles or multiple nodules on one image. Accordingly, a selection of the generated images was included for each patient. More specifically, the selection process is that if the images represent the same nodule, those images would have the same diagnostic label. If the images from the same patient are showing different nodules, the definitive diagnostic label was assigned based on the corresponding histopathology or cytology findings. Images without clear cytology or histopathological results were excluded. As a result, this study involves 917 ultrasound images for investigations.

This research incorporated the benchmark dataset for comparison to evaluate the CAD models. The Digital Database Thyroid Image (DDTI), an open-access ultrasound images database offered by Pedraza et al. [243], was utilised. In order to align with existing studies [290, 221], we have selected our benign images with TIRADS scores at 1, 2, and 3, while the remaining scores $4a$, $4b$, $4c$, and 5 were labelled as malignant. As a result, a

total of 448 open-access thyroid ultrasound images were selected, with 66 benign and 382 malignant images.

CT Dataset

CT scans are efficient in detecting abnormal thyroid structures based on the shapes and densities [279]. From the 578 patients, this research sliced the entire CT volumes with $5mm$ spacing for each patient, which allows the selection of images with distinct structures for CAD training to avoid over-fitting and bias. After segmenting the sliced CT scans into left and right sides in the middle of the trachea, the representative CT slice was selected and assigned labels correspondingly referred to the diagnostic reports. As a result, $2,257$ CT scans were involved in this study. The distribution of the ultrasound and CT scans in the six classes is presented in Table 5.2. Figure 5.5 displays some sampled images of the acquired datasets in the six classes.

Table 5.2: Distribution of the datasets in the six classes.

Modality		Normal	Thyroiditis	Cystic	Goitre	Adenoma	Cancer	**Total**
Ultra	CN	15	80	396	167	59	200	917
CT	LT	253	95	357	180	86	209	1180
	RT	246	49	394	187	68	133	1077

5.4.2 Data Imbalance

In practice, data imbalance issue is considered a common issue for computer vision tasks [129]. Data augmentation techniques, such as rotating, flipping, random cropping, stretching, or mirroring are usually applied for natural images [358]. However, those techniques are not appropriate in this case. This research focuses on interpreting the textures, structures, and especially position features of the thyroid images. Thus, altering the information from the original images is not suitable.

In order to address the data imbalance issue, this study utilises the stratified cross-validation (CV) [251, 195] and the categorical cross-entropy (CCE) techniques. The strat-

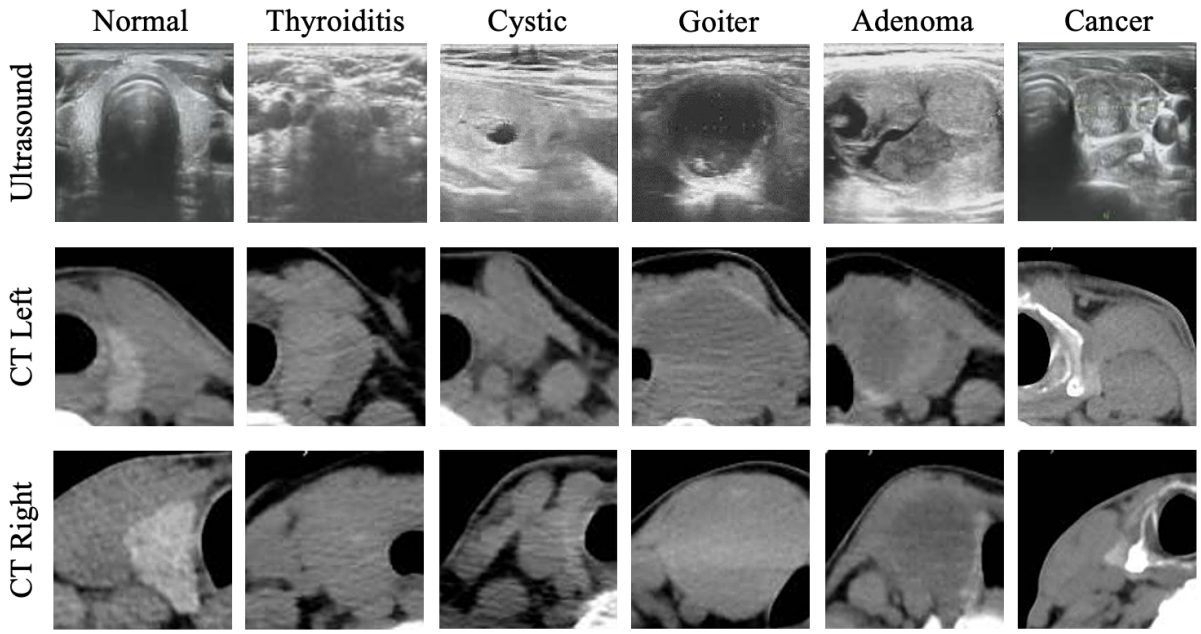


Figure 5.5: Sample images of ultrasound and CT for the six classes.

ified CV is a practical training and testing split technique, which divides the dataset in each iteration to retain the distribution of the original observations for each class. This technique compensates for the unequal number of classes and the uneven distribution among the classifications. Moreover, it allows the variance of the estimates in each fold to be reduced, enhancing the fairness of the generated results [251]. Therefore, the stratified CV was applied in this case rather than the standard CV split so that the lop-sidedness of each class can be maintained so that the implementation of the CAD can be much more accurate and efficient without performing data augmentation pre-processing.

The second approach to address data imbalance is using categorical cross-entropy (CCE) as the loss function to reach unbiased results. The CCE is usually applied to assign weights to different classes as it can adapt the penalty of a probabilistic false-negative rate for a given class [108]. The CCE is calculated by using Eq. 5.3. With the encoded labels y, the last fully connected layer would produce a feature vector indicating the possibility for each class. Here, s_p denotes the predicted score for the specific class, s_j is the inferred score for each class in C, and C refers to the total number of classes.

$$CCE = -log(\frac{e^{s_p}}{\sum_j^C e^{s_j}}) \tag{5.3}$$

5.4.3 Parameter Settings

The classification results were calculated with the stratified 10-fold CV during each iteration. More specifically, the best-performing epoch was selected for each fold. The final result was calculated by averaging the ten best-performing epoch scores. The overall implementation process can be viewed in Algorithm 3. In this case, accuracy, precision, recall, specificity, NPV, and F1 scores were incorporated and calculated using Eqs. 5.4 to 5.9 from the generated confusion matrix. K is the total number of folds, TP as "True Positive", TN as "True Negative", FP as "False Positive", and FN as "False Negative".

$$Accuracy = \frac{1}{K}\sum\nolimits_{i=1}^{K}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{5.4}$$

$$Precision\ (PPV) = \frac{1}{K}\sum\nolimits_{i=1}^{K}\frac{TP_i}{TP_i + FP_i} \tag{5.5}$$

$$Recall\ (Sensitivity) = \frac{1}{K}\sum\nolimits_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \tag{5.6}$$

$$Specificity = \frac{1}{K}\sum\nolimits_{i=1}^{K}\frac{TN_i}{TN_i + FP_i} \tag{5.7}$$

$$NPV = \frac{1}{K}\sum\nolimits_{i=1}^{K}\frac{TN_i}{TN_i + FN_i} \tag{5.8}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.9}$$

During the implementation stage, the Adam optimiser was applied with an initial learning rate of 1×10^{-2}, and it was gradually updated through the gradient descent

algorithm in the fine-tuning stage. The learning rate reached 1×10^{-5} showing stability and was set fixed thereafter. The batch size was set to 10 during the training process.

In order to reach consistency for comparison, all the experiments were performed under the same computational environment on the Tensorflow platform, with a 64-bit Windows 10 Pro desktop, which had an Intel Core i7-9700 processor with 16 gigabytes of memory and a GeForce GTX 1050 GPU.

Algorithm 3: Pseudo-code for image classification with CNN through 10-fold stratified cross-validation

Input: $X = \{X_1, X_2, \ldots, X_n\}$; X is labeled image set and n is total number of instances
$y \in \{0, 1\}$ or $y \in \{0, 1, 2, 3, 4, 5\}$; y is the class labels
Divide data into stratified k folds
Initialisation:
Set $i = 0$
while $i \leq$ *Iteration* **do**
  $i = i + 1$
  for k_i *in* K *folds* **do**
    Set fold k_i as testing set
    Train CNN to extract feature vectors from remaining $K - k_i$ folds
    $CNN \leftarrow (X, y)$; Training image sets and labels will be sent to CNN
    for $i \in \{1 : len(y)\}$ **do**
      $\mathcal{P}^{(i)} \leftarrow F^{(i)}$; Predict the test image class based on extracted features using Eq. 5.3
      Output: $\mathcal{P} = \mathcal{P}^{(i)}, \mathcal{P}^{(i+1)}, .., \mathcal{P}^{(i+m)}$; Set of testing image class labels
    end
    Calculate correctly classified image in fold k_i using Eq. 5.4
  end
  $Acc = Acc^{(1)}, Acc^{(2)}, .., Acc^{(i)}$; Accumulate the accuracy scores for each iteration and store
end
Calculate average performance of all K folds
Output: Averaged testing accuracy for K folds $\rightarrow \overline{Acc}$

5.5 Results

This research is the first of its kind, which adopts two pre-operative medical image modalities to diagnose thyroid disease. A group of experiments was conducted to compare the CAD models between ultrasound and CT images. With the labelled datasets, the

binary and multi-class classification tasks were performed independently. Their results are demonstrated, explained, and compared in the following sections.

5.5.1 Performance of Binary Classification

In order to have an explicit demonstration of the performance comparison between the ultrasound and the CT images, this section interprets the binary classification results with the selected CNNs.

Ultrasound Performance - Binary

The two sources of ultrasound images were used to evaluate the selected 11 models through the binary classification task. All the models were evaluated through accuracy, precision, NPV, recall, F1 scores, number of parameters, and running time in minutes.

Table 5.3 presents the experimental results for the binary classification task with the two sets of ultrasound images. For both data sources, Xception reached the best averaged-accuracy rates of 0.980 for the DDTI dataset and 0.987 for the open-access dataset. The second best-performing model for the DDTI dataset was DenseNet121 with an accuracy of 0.978. Similarly, DenseNet121 was also the second best-performing model for the CN dataset with an accuracy of 0.965.

Regarding the running time comparison, InceptionResNetV2 is considered the most time-consuming model for both datasets. ResNet10 is the most efficient model for both datasets, with 33 minutes for the 10-fold CV using the DDTI dataset, and it took 46 minutes to process the CN dataset. Xception reached a similar running time with InceptionV3 and DenseNet121. Therefore, those three models were further selected for the multi-classification task due to the promising accuracy and efficiency reached.

CT Performance - Binary

Since there is no benchmark dataset for CT images to be compared, this research only used the CN dataset to evaluate the 11 CNN models. Table 5.4 presents the binary

Table 5.3: Binary classification results for ultrasound images.

Model	ACC	PPV	NPV	Recall	F1	No. para	min
DDTI Results							
VGG8	0.857	0.830	0.667	0.860	0.845	5,516,610	46
VGG11	0.832	0.830	0.156	0.635	0.720	10,826,306	58
VGG16	0.853	0.730	0.136	0.850	0.785	16,320,514	118
VGG19	0.783	0.853	0.156	0.901	0.876	21,630,210	141
ResNet10	0.864	0.880	1.000	0.860	0.870	4,912,578	33
ResNet18	0.873	0.870	0.846	0.870	0.870	11,187,138	53
ResNet50	0.850	0.730	0.852	0.850	0.785	23,591,810	82
DenseNet121	0.978	0.985	1.000	0.925	0.954	7,039,554	149
Xception	**0.980**	**0.990**	**1.000**	**0.945**	**0.967**	20,865,578	131
InceptionV3	0.967	0.980	1.000	0.885	0.930	21,806,882	117
InceptionResNet	0.971	0.985	1.000	0.900	0.941	54,339,810	182
CN Data Results							
VGG8	0.799	0.790	0.802	0.800	0.795	5,516,610	94
VGG11	0.815	0.688	0.827	0.275	0.393	10,826,306	142
VGG16	0.809	0.800	0.814	0.810	0.805	16,320,514	241
VGG19	0.792	0.680	0.795	0.090	0.151	21,630,210	297
ResNet10	0.832	0.830	0.828	0.830	0.830	4,912,578	46
ResNet18	0.828	0.830	0.828	0.830	0.830	11,187,138	109
ResNet50	0.865	0.870	0.861	0.860	0.865	23,591,810	218
DenseNet121	0.965	0.965	0.965	0.930	0.947	7,039,554	203
Xception	**0.987**	**0.985**	**0.990**	**0.975**	**0.980**	20,865,578	214
InceptionV3	0.924	0.925	0.928	0.840	0.880	21,806,882	241
InceptionResNet	0.957	0.982	0.952	0.820	0.894	54,339,810	377

classification results with CT images. Unsurprisingly, Xception again outperformed all the other models, with an accuracy of 0.966 for the left-side CT scans and 0.970 for the right-side CT scans. DenseNet121 was still the best-performing model with an accuracy of 0.954 for the left CT and 0.940 for the right CT images.

Table 5.4: Binary classification results for CT images.

Model	ACC	PPV	NPV	Recall	F1	No. para	min
CT Left Results							
VGG8	0.801	0.807	0.659	0.984	0.887	5,516,610	130
VGG11	0.687	0.807	0.268	0.794	0.800	10,826,306	159
VGG16	0.798	0.800	0.727	0.994	0.886	16,320,514	275
VGG19	0.678	0.796	0.233	0.796	0.796	21,630,210	321
ResNet10	0.815	0.845	0.603	0.938	0.889	4,912,578	52
ResNet18	0.812	0.828	0.634	0.962	0.890	11,187,138	81
ResNet50	0.895	0.902	0.860	0.974	0.936	23,591,810	221
DenseNet121	0.954	0.955	0.953	0.989	0.972	7,039,554	281
Xception	**0.966**	**0.961**	**0.986**	**0.997**	**0.979**	20,865,578	263
InceptionV3	0.914	0.927	0.855	0.968	0.947	21,806,882	207
InceptionResNet	0.892	0.906	0.821	0.964	0.934	54,339,810	510
CT Right Results							
VGG8	0.785	0.788	0.707	0.986	0.876	5,516,610	106
VGG11	0.738	0.802	0.389	0.877	0.838	10,826,306	143
VGG16	0.669	0.787	0.277	0.784	0.786	16,320,514	270
VGG19	0.777	0.780	0.667	0.992	0.873	21,630,210	297
ResNet10	0.774	0.822	0.506	0.904	0.861	4,912,578	48
ResNet18	0.780	0.810	0.537	0.935	0.868	11,187,138	74
ResNet50	0.878	0.886	0.832	0.966	0.924	23,591,810	203
DenseNet121	0.940	0.938	0.948	0.987	0.962	7,039,554	260
Xception	**0.970**	**0.967**	**0.983**	**0.995**	**0.981**	20,865,578	240
InceptionV3	0.895	0.914	0.817	0.954	0.934	21,806,882	205
InceptionResNet	0.872	0.877	0.847	0.971	0.922	54,339,810	417

A similar pattern was again demonstrated where ResNet10 is the most efficient model, and InceptionResNetV2 took the most running time. Figure 5.6 illustrates the detailed running time comparison of the models.

5.5.2 Performance of Multi-Classification

There is no benchmark dataset for the multi-class classification task for either imaging modality. Therefore, the CN dataset was used to evaluate the three selected models,

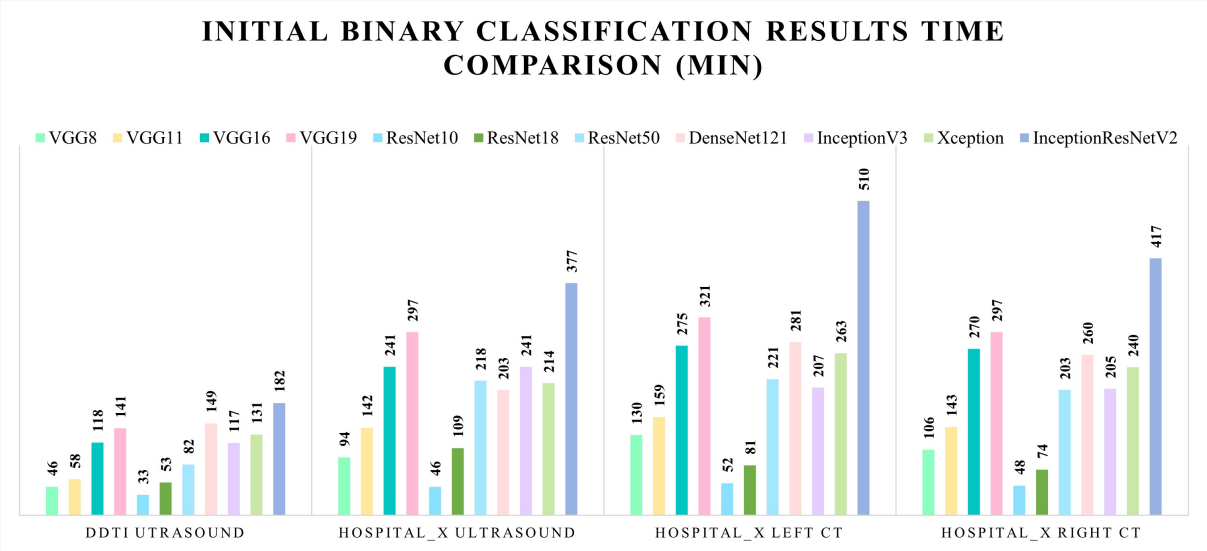


Figure 5.6: Running time comparison of the 11 CNN models.

including InceptionV3, DenseNet121, and Xception, as they are the most accurate and efficient models in the binary classification task.

Ultrasound Performance - Multi-class

Table 5.5 presents the multi-class classification results with the CN ultrasound image set. Xception generated the highest accuracy compared to InceptionV3 and DenseNet121, with an F1 score of 0.93, 0.95, 0.99, 0.96, 0.90, and 0.98 for the normal, thyroiditis, cystic nodule, multi-nodular goitre, adenoma, and cancer classes, respectively. The averaged accuracy score through the 10-fold stratified CV was 0.97, 0.85 and 0.89 for Xception, InceptionV3 and DenseNet121, correspondingly.

Table 5.5: Multi-class classification results for ultrasound images.

Class	Precision			Recall			F1 Score		
	Incep.	DN	Xcep.	Incep.	DN	Xcep.	Incep.	DN	Xcep.
Normal	0.70	0.74	**1.00**	0.50	0.60	**0.87**	0.58	0.66	**0.93**
Thyroiditis	0.84	0.90	**0.91**	0.70	0.66	**1.00**	0.76	0.76	**0.95**
Cystic	0.84	0.90	**0.99**	0.95	0.98	**0.99**	0.89	0.94	**0.99**
Goitre	0.86	0.84	**0.96**	0.81	0.90	**0.97**	0.83	0.87	**0.96**
Adenoma	0.74	**0.90**	**0.90**	0.53	0.63	**0.90**	0.61	0.74	**0.90**
Cancer	0.89	0.91	**1.00**	0.89	0.93	**0.96**	0.89	0.92	**0.98**

Figure 5.7 shows the confusion matrix of the Xception model on the ultrasound images. The Xception model correctly classified all normal images and thyroid cancer images. The model also correctly classified 90.91% of the images of thyroiditis, 98.74% of cystic, 95.86% of multi-nodular goitre, and 89.83% of adenoma.

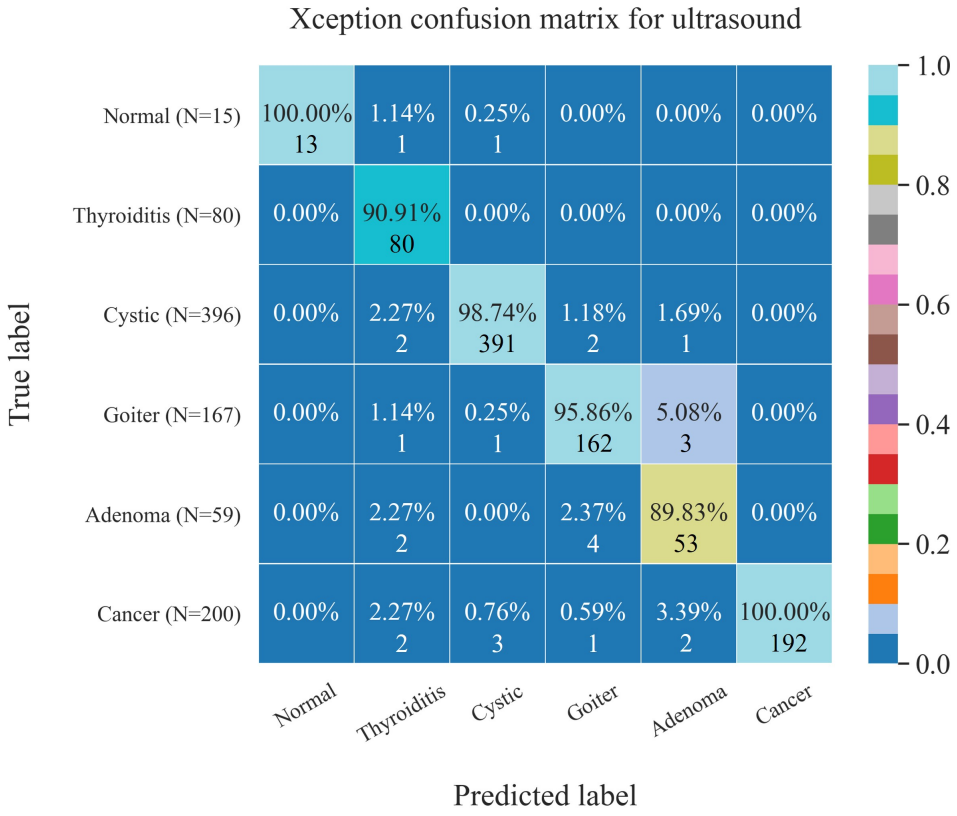


Figure 5.7: Confusion matrix of the multi-class classification task on ultrasound.

CT Performance - Multi-class

Table 5.6 presents the classification results for the segmented left-side and right-side CT images separately. For the left-side CT, the accuracy rates were 0.95, 0.68 and 0.79 for Xception, InceptionV3 and DenseNet121 respectively. For the right-side CT, the corresponding accuracy rates were 0.94, 0.67, and 0.79. Xception generated F1 scores of around 0.95 for both sides of the gland in the six classes.

Figure 5.8 depicts the confusion matrix for both sides of the CT images. For the left-side thyroid lobe, Xception reached a precision of 96.79% of the normal class, 94.95%,

Table 5.6: Multi-class classification results for CT images.

Class	Precision			Recall			F1 Score		
	Incep.	DN	Xcep.	Incep.	DN	Xcep.	Incep.	DN	Xcep.
CT Left									
Normal	0.76	0.86	**0.97**	0.79	0.88	**0.95**	0.78	0.87	**0.96**
Thyroiditis	0.67	0.84	**0.95**	0.69	0.66	**0.99**	0.68	0.74	**0.97**
Cystic	0.61	0.73	**0.94**	0.90	0.97	**0.96**	0.73	0.83	**0.95**
Goitre	0.73	0.79	**0.95**	0.70	0.77	**0.92**	0.71	0.78	**0.94**
Adenoma	0.56	0.83	**0.94**	0.62	0.67	**0.94**	0.59	0.74	**0.94**
Cancer	0.68	0.83	**0.93**	0.71	0.79	**0.92**	0.70	0.81	**0.93**
CT Right									
Normal	0.64	0.84	**0.89**	0.87	0.89	**0.97**	0.74	0.86	**0.93**
Thyroiditis	0.53	0.69	**1.00**	0.50	0.63	**0.90**	0.51	0.66	**0.95**
Cystic	0.71	0.78	**0.96**	0.93	0.97	**0.95**	0.80	0.86	**0.95**
Goitre	0.62	0.77	**0.97**	0.72	0.80	**0.91**	0.67	0.78	**0.94**
Adenoma	0.60	0.75	**1.00**	0.58	0.71	**0.88**	0.59	0.73	**0.94**
Cancer	0.67	0.78	**0.91**	0.66	0.77	**0.95**	0.68	0.78	**0.93**

93.68%, 94.86%, 94.19%, and 93.24% of the thyroiditis, cystic, goitre, adenoma, and cancer types, respectively. For the right-side CT, Xception correctly classified all the thyroiditis and adenoma types and achieved 88.81%, 95.90%, 96.59%, and 91.37% accuracy of normal, cystic, goitre, and cancer types.

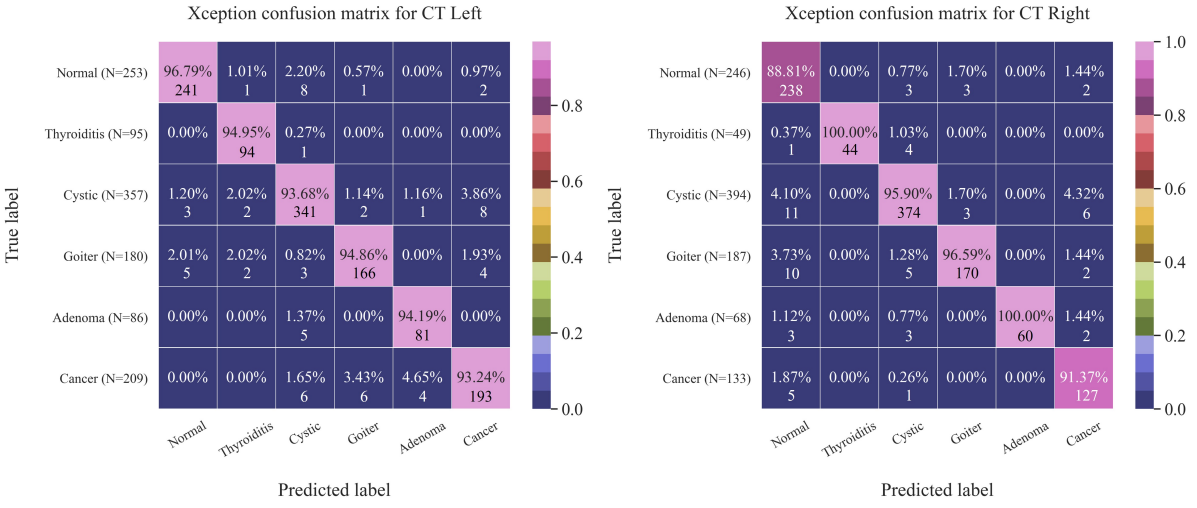


Figure 5.8: Confusion matrix of the multi-class classification task on CT.

Figure 5.9 illustrates the averaged F1 scores of the 10-fold stratified CV approach in the six classes for both imaging modalities. Evidently, Xception architecture was much more

stable and robust in classifying images among the six classes compared to InceptionV3 and DenseNet121.

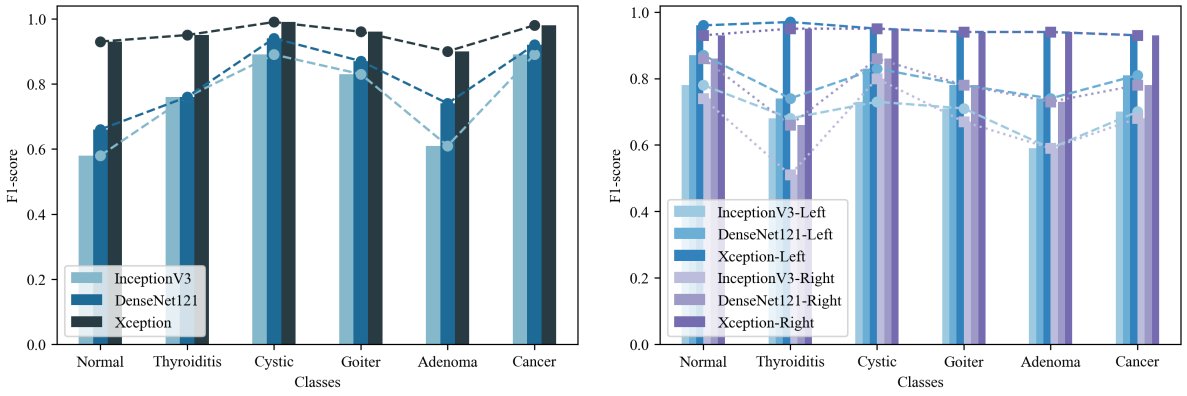


Figure 5.9: Averaged CV F1 scores for the CNNs on ultrasound and CT (left to right).

Xception provided the most accurate diagnostic rates among all the models in both binary and multi-class classification tasks. Moreover, it also outperformed the other two architectures (i.e., InceptionV3 and DenseNet121) with the least running time required for training and testing. Figure 5.10 presents the running time comparison of the three models where Xception finished training and testing ultrasound, left-side CT, and right-side CT scans in 106, 138, and 127 minutes, respectively.

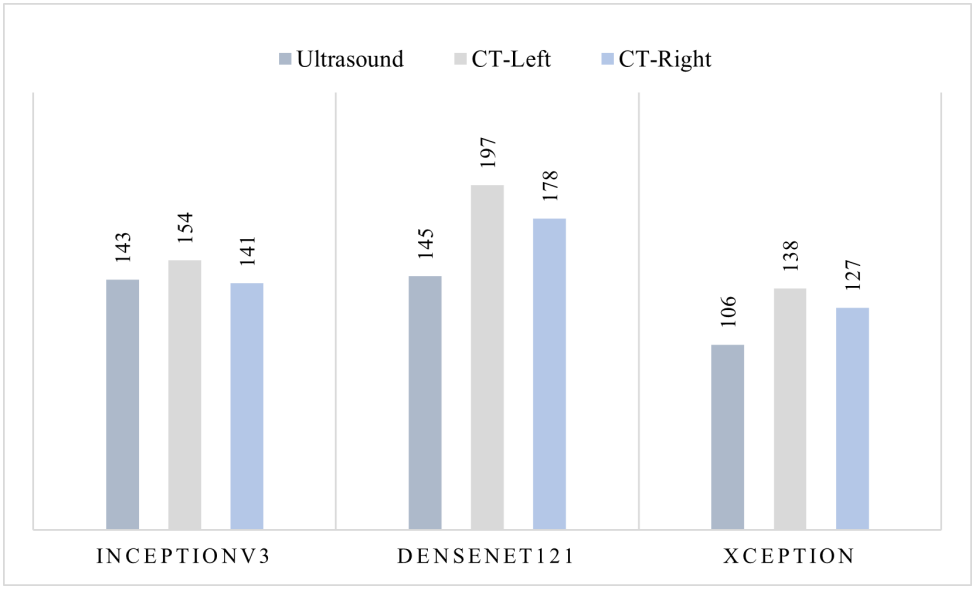


Figure 5.10: Running time comparison for the CNNs on both image modalities.

5.6 Discussion

Automatic thyroid disease diagnosis can be achieved by incorporating deep learning techniques and medical images, potentially mitigating patients' financial and physical pressure from clinical diagnosis. The deep learning-driven diagnosis can also reduce human false-positive and false-negative rates when making decisions. Extensive experiments were conducted with the benchmark and real-world datasets to evaluate the selected CNN models. This section interprets the classification results and highlights the comparison between the selected CNNs, the two image modalities, and the performed tasks.

5.6.1 Binary Classification Discussion

The existing studies incorporating thyroid images with deep learning techniques can reach an accuracy of around 80% [188, 146, 95, 90, 53]. Most works in the literature were conducted on a few CNN models using unitary image modality, which did not ensure a comprehensive analysis of the existing architectures. Therefore, this research adopted 11 CNN models for thyroid disease detection on two pre-operative medical image modalities for comparison.

More specifically, the reduced number of layers for VGG and ResNet models increased the diagnostic accuracy, indicating that the shallower layer CNN architectures work well on the simple structure of images. The selected Xception model has obtained a classification accuracy of 0.980 using the DDTI dataset and 0.987 using the CN ultrasound images. The accuracy rates of 0.966 for the left-side CT and 0.970 for the right-side CT scans were generated with Xception. Both image modalities outperformed the existing studies. Unsurprisingly, ultrasound images performed slightly better on the task than CT scans. The potential reason behind this might be related to the characteristics of ultrasonography, as it is sensitive to human intervention. During the image acquisition process, clinicians tend to screen nodules that have apparent features for diagnosing, which might contribute to increased classification accuracy. Whereas the CT scan is generated fully

automatically, resulting in the lower classification results. However, CT scans can reach comparable performance with ultrasound images. This outcome also highlights that CT scans can be of potential use for the implementation of CAD designs.

5.6.2 Multi-class Classification Discussion

The existing studies on multi-classifying thyroid disease subtypes are considerably lacking, and this research bridges the literature gap and interprets the findings.

The experimental results showed that Xception produced superior accuracy and was also efficient in processing images, thus allowing the input of a more vast number of images, making it suitable for the clinical domain. The correct classifications of the ultrasound images with thyroid cancer (i.e., 100%) are superior to those recently reported by Chi et al. [55] and Liang et al. [166]. Xception was also highly accurate in classifying ultrasound images with thyroid cysts or normal thyroid glands. The lower accuracy of classifying thyroiditis, multi-nodular goitre, and adenoma images was likely due to these entities' overlapping characteristics on the ultrasound images.

The classification of the cancer images was somewhat lower for CT than for ultrasound. This, again, was unsurprising as interior features and characteristics of thyroid cancer were better demonstrated on ultrasound. In contrast, CT may be more suitable for demonstrating the characteristics external to the thyroid gland, such as invasion or erosion of the adjacent structures. Another potential contributing factor was how the images were acquired since the operator-dependent test might select images containing more characteristic appearances of each pathology class.

In thyroidology, the existing diagnostic workflow has been well-established over decades. Any new technology that may significantly disrupt the current status would first need to earn the confidence of clinicians and patients. Clear demonstration of accuracy, efficiency, and ease of use are some of the foundations required to develop that confidence. To address the issue and prove the feasibility of the CAD techniques, this research proposed applying different medical image modalities to reach a comprehensive level of diagnostic

decision-making. However, this research has limitations regarding the labelling process. In practice, one or several pathology diagnoses can co-exist within one side of the thyroid gland, whereas this study assigned labels to each image based on the dominating class. The future plan would incorporate multi-labels into one image for precision classification.

5.7 Summary

In order to elevate thyroid disease diagnostic accuracy and efficiency, the proposed CNN-BM framework was used to evaluate two pre-operative medical imaging modalities through binary and multi-class classification tasks.

11 different CNN architectures were utilised to distinguish between normal and abnormal thyroid nodule images from ultrasound and CT images. To benchmark the performance, the open-access DDTI dataset was incorporated. Xception demonstrated the best performance on both binary and multi-class classification tasks through both image modalities. It reached an accuracy of around 0.98 for the ultrasound and 0.97 for the CT scans. The comprehensive diagnosis was further achieved through a multi-class classification task. The three best-performing (i.e., models with the highest accuracy and efficiency) CNN architectures were utilised in this case. The averaged accuracy scores for ultrasound and CT images were 0.972 and 0.942 for the six classes.

In summary, this comparative research illustrates that Xception can be adaptive to different image modalities with superior performance. Ultrasound images generate better diagnostic accuracy rates than CT scans when establishing CAD models. Such CAD systems release patients' burdens from the clinical diagnostic process and emulate the expert-level diagnosis to assist clinicians with offering preliminary decisions. The proposed CAD systems can be further adapted to different diseases, enhancing clinical applications.

Chapter 6

Multi-Channel Deep Convolutional Neural Network Architectures in Thyroid Disease Detection

6.1 Introduction

In recent years, deep CNN models have yielded unprecedented performance on thyroid cancer diagnosis through diversified imaging modalities [189, 2, 23, 38]. Such deep learning-based CAD systems tend to provide superior diagnostic accuracy and efficiency, sometimes even outperforming clinicians [345]. Although the application of such CAD systems is commonly adopted, there is an ever-growing demand for more advanced models to address more challenging scenarios.

Functional thyroid disease, such as thyroiditis, can co-exist with neoplastic thyroid disease, like adenoma or cancer, and so can the same category [169]. However, the existing studies have considerably ignored the phenomenon where an individual patient might suffer from various types of thyroid diseases at one time. The majority of the existing CAD systems can only distinguish between benign and malignant thyroid nodules individually. These systems are efficient in detecting the disease and cannot offer a comprehensive

diagnosis for the entire gland. Therefore, those CAD systems cannot be widely approbated in the clinical domain as it has limitations in providing an expert-level diagnosis.

Accordingly, this research study proposes the multi-channel CNN architectures to reach expert-level diagnosis from three aspects, including 1) elevating the comprehensiveness of diagnostic decision-making, 2) reaching patient-specific design for CAD implementations, and 3) enhancing the interpretability of the generated decision. More specifically, three multi-channel architectures were introduced and can be adaptive based on different diagnostic output choices, denoting the status of the entire gland. These architectures ensure that the diagnoses of the overall gland are made for the individual patient at a time. The generated diagnosis can easily be comprehended by clinicians, emphasising the interpretability of the produced results by such "black-box" approaches. The architectures were majorly designed following the characteristics of CT scans and were further evaluated through different gender groups to evaluate their generalisation.

Therefore, this research study proposes three adaptive multi-channel CNN architectures to help streamline the diagnostic process for thyroid disease, highlighting the disease subtype co-existence phenomenon and aiming to establish an accurate CAD system following the patient-specific design. The proposed multi-channel approach demonstrates enhanced diagnostic accuracy and has the potential to be integrated into the clinical workflow to guide primary care physicians in deciding if a specialist referral is warranted. The related background works of this chapter were described in Chapter 2.3.2.

The contents of this chapter have been published in journal articles 1, 3.

6.2 Problem Formulation

Given the image set $\mathcal{D} \in \{X, y\}$, the image matrix $X = \{X_1, X_2, X_i, \ldots, X_n\}$ and label $y \in \{0, 1, 2, \ldots, 5\}$ follow the same protocol as the multi-class classification task described in Chapter 5.2. The proposed CAD system provides a comprehensive diagnosis of the entire thyroid gland.

This chapter proposes the multi-channel CNN architectures where they are inspired by the idea that kernel size has a considerable impact on CNN performance [162, 112, 330]. Kernel size convolves feature maps by determining the size of the receptive fields. More specifically, a larger kernel will generate abstract features, and a smaller kernel will learn more detailed textures from the input image. Şaban Öztürk et al. [363] once deployed histopathology images for evaluation and reported a moderately strong association between the kernel size and the CNN performance. Several other studies accentuated that multi-channel CNN architectures generate enhanced diagnostic performance [171, 297, 11].

On top of elevating the diagnostic accuracy by concatenating feature maps generated through diversified kernel sizes, the proposed multi-channel CNN models have three adaptive architectures tailored for distinct diagnostic scenarios. Besides, those architectures emphasised the patient-specific design for CAD systems and the detection of thyroid disease subtype co-existence phenomenon.

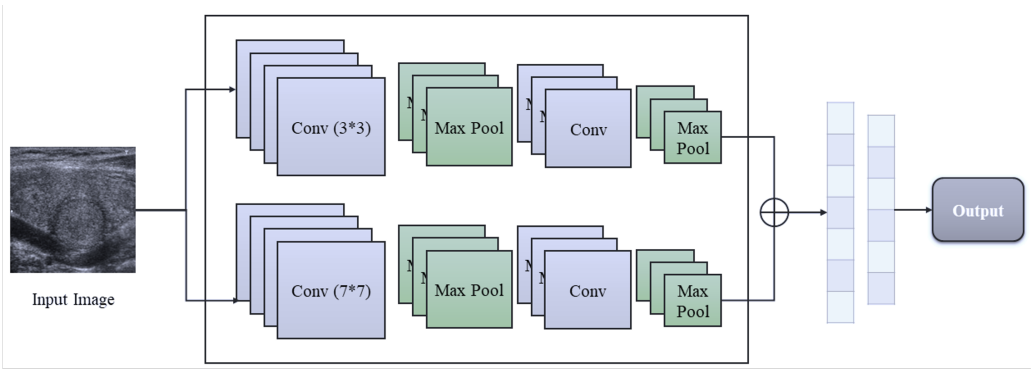
6.3 Methodology

This thesis proposed three adaptive multi-channel CNN architectures tailored for diversified use of results generation for thyroid disease diagnosis. This section describes the three architectures, highlighting their benefits and applications.

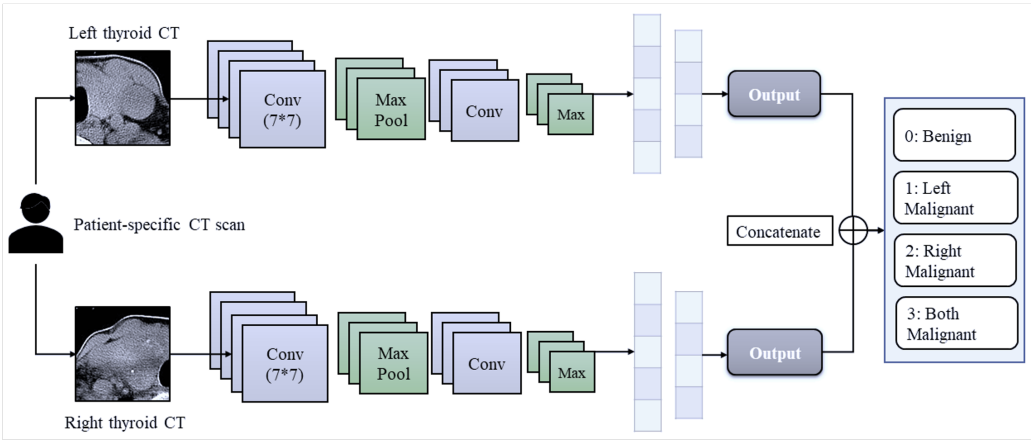
6.3.1 The Proposed MTCD Framework

Figure 6.1 illustrates the proposed multi-channel CNN framework for thyroid disease detection (MC-TDD), consisting of the single input dual-channel (SIDC), double inputs dual-channel (DIDC), and four-channel architectures.

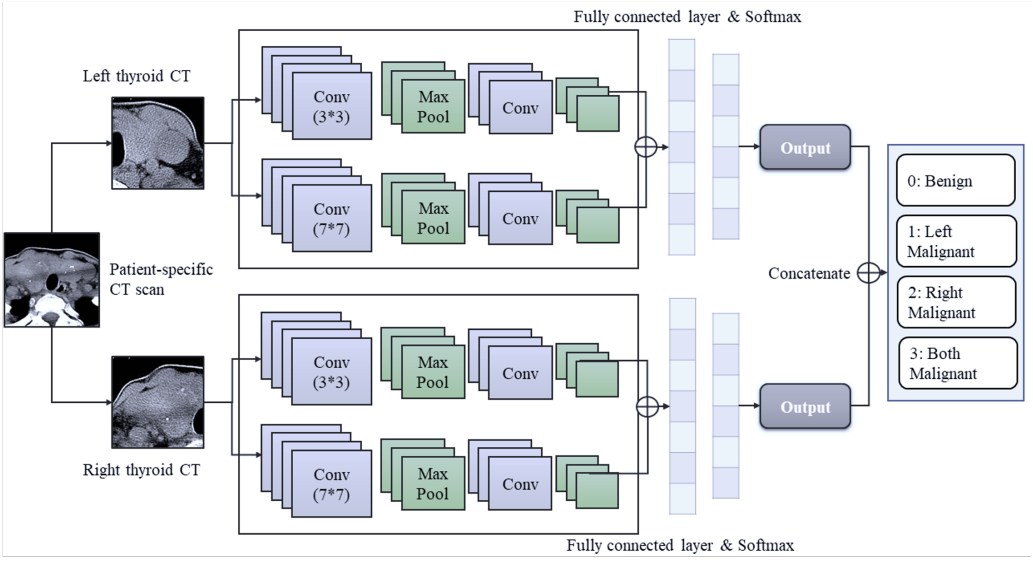
The conventional CNN models usually apply a fixed kernel size for convolutional operations. This limits the comprehensiveness of learnt features as information loss gradually aggravates during each convolutional operation. The motivation behind the multi-channel architectures is to mitigate the risk of losing too much critical information. Accordingly,



a. Single Input Dual-Channel (SIDC)



b. Double Inputs Dual-Channel (DIDC)



c. Four-Channel

Figure 6.1: Multi-channel Thyroid Disease Detection (MC-TDD) framework.

the proposed model obtains an intermediate feature map by concatenating the outputs generated through convolutions from different kernel sizes so that more informative features can be learnt from the original image.

The commonly used kernel sizes in the computer vision tasks are 1×1, 3×3, 5×5, 7×7, and 9×9 [293, 363]. The comparative study by Şaban Öztürk et al. [363] indicated that CNN tended to have the highest validation errors with a kernel size of 9×9. Thus, the 9×9 kernel size was excluded from this research as it could not learn sufficient features for making classifications. Accordingly, this research considers 1×1 and 3×3 kernels as "*smaller kernel sizes*", and 5×5 and 7×7 as "*larger kernel sizes*". By concatenating the feature maps generated from a smaller and a larger kernel sizes, the enumeration of those sizes was evaluated, respectively.

After deciding the choice of the kernel sizes, the number of convolutional channels is also critical. In theory, more channel numbers would have better classification performance as more features can be obtained. At the same time, the model is also prone to over-fitting as it has a vast number of parameters [328]. In this regard, the dual-channel architecture is more appropriate than three or more channel numbers as it requires less computational resources to train, increases diagnostic accuracy, and is suitable for medical images with simple structures.

6.3.2 Multi-channel CNN Architectures

With the objective of enhancing thyroid disease diagnostic accuracy, this section describes the proposed three types of multi-channel CNN architectures.

Single Input Dual-Channel (SIDC)

Given the input dataset $\mathcal{D}$, each encoded image X_i will travel through two convolutional channels simultaneously with different kernel sizes (i.e., one with a larger size and another one with a smaller size), then following the max-pooling operations and produce two feature maps. The produced feature maps will then be concatenated to the fully

connected layer and generate the classification output, denoting the class of the input image. The main advantage of the SIDC architecture is as follows:

- The intermediate feature map from two kernel sizes can maintain more spatial information from the original image [112], eventually increasing the classification accuracy.

Although the SIDC architecture enhances diagnostic accuracy, yet it can only make decisions for one image at a time. By looping through the dual-channel architecture, different kernel sizes can automatically learn different features from the input images. In other words, larger kernel sizes will learn more abstract features, while smaller kernel sizes can learn detailed textures such as edges. The concatenating of the two feature maps following a fully connected layer with a softmax operator will be used to interpret the selected features [222].

Double Inputs Dual-Channel (DIDC)

Although the SIDC architecture is dedicated to elevating diagnostic accuracy, the model detects abnormal thyroid lesions based on the individual input image and cannot produce a comprehensive diagnosis for the entire thyroid gland for each patient. The proposed double inputs dual-channel (DIDC) model addressed this limitation and achieved the patient-specific diagnosis.

Notably, the DIDC architecture is designed following the characteristics of CT scans. Protocolised acquisition of CT scans can provide consistent quality images and a complete view of the thyroid gland and its surrounding structures. Segmenting the overall CT scan allows a different label to be applied to each side. The workflow of the proposed multi-channel CNN is presented in Figure 6.2.

The process is that a thyroid CT volume will be sliced, and representative slices will be chosen following the $5mm$ spacing. The selected CT scans from an individual patient will be segmented into the left and right sides, and both sides will need to be labelled separately. The second step is to apply the 10-fold stratified CV for training and testing

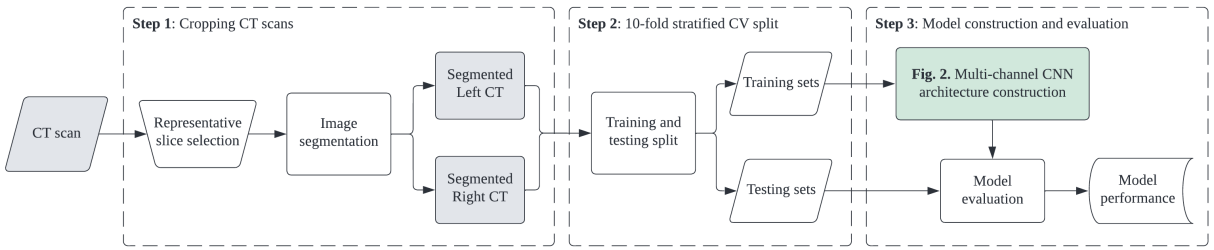


Figure 6.2: Multi-channel CNN model implementation procedure.

splits. In the third step, the segmented CT scans from the training sets will be fed into the constructed multi-channel CNN model, while the testing sets will be used to evaluate the model. Each left and right-side CT scan will travel through a single-channel CNN model with the same kernel size to obtain receptive fields with the same size. Then, the outputs of the two channels will produce a diagnostic result for each side of the image. Furthermore, the results from the two channels will be concatenated into a 4×4 matrix, denoting the status of the entire thyroid gland. More specifically, the DIDC model classifies the patient into "*benign*" represented using 0, "*left-side malignant*" denoted using 1, "*right-side malignant*" using 2, and "*both sides malignant*" represented using 3. The proposed DIDC model can mimic real-life clinical diagnoses made for an overall thyroid gland, reaching the patient-specific design for diagnosis. The proposed DIDC model has the following comparative advantages:

- It reaches the patient-specific design for the diagnosis of thyroid disease.
- It offers the explicit status of both sides of the gland for a comprehensive diagnosis.
- The architecture outputs the overall status of the gland, making diagnostic and treatment decisions more efficiently.

Since this model follows the patient-specific diagnostic protocol, some prerequisites must comply. The segmented left and right sides of the CT images must be applied in two separate convolutional streams so the model can provide a diagnostic decision for each side simultaneously. The input streams of the left and right sides must be from

one patient. Additionally, the scale of the two input streams (i.e., the number of both input image sets) must be equal. Similarly, these requirements also apply to the proposed four-channel architecture.

Four-Channel

Beyond the SIDC and DIDC architectures, the four-channel architecture incorporates enhanced diagnostic accuracy and patient-specific design benefits. Each of the left and right sides of the CT scans will be processed with dual-channel convolutions under different kernel sizes. Subsequently, the processed feature maps will be concatenated and generate a classification output that indicates the diagnosis for the entire thyroid gland. The advantages of the four-channel architecture are as follows:

- It persists the increased accuracy rates obtained by the SIDC model since two different kernel sizes were incorporated to generate the intermediate feature map.
- It is also designed as patient-specific obtained from the DIDC architecture.
- It enhances the diagnostic accuracy, in the meantime, provides a comprehensive diagnosis for the overall gland more accurately and effectively.

In order to detect thyroid disease co-existence, the four-channel architecture is incorporated for a multi-class classification task. Figure 6.3 depicts the four-channel model for multi-classifying thyroid disease.

This study incorporates the six most commonly seen types for evaluation (refer to Chapter 5). The generated classification vectors denoting the class of both left and right sides will be further concatenated into a 16×16 matrix, indicating the overall status of the gland. The 16 classes include the six pre-defined classes and ten combinations of disease types in each lobe (see Figure 6.3). The specific fusion of the diagnosis was made when falling into the below scenarios:

- If both sides are in the normal status, the final class would be the "*Normal*" type (class 0 in Figure 6.3).

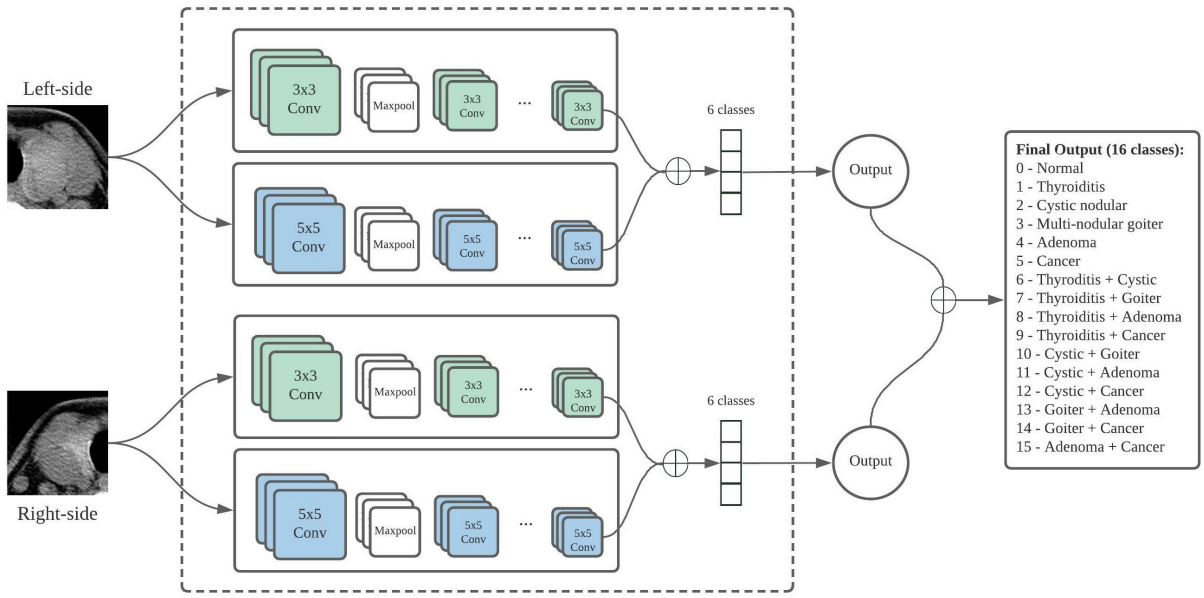


Figure 6.3: Four-channel architecture for thyroid disease co-existence detection.

- If one side of the lobe is normal and the other side is with a particular disease, or when both sides have the same type of disease, the final class would be the disease type (from class 1 to 5 in Figure 6.3).
- If both sides appear to have different types of diseases, the final class would be the fused type (from classes 6 to 15 in Figure 6.3).

The pseudo-code for implementing the four-channel architectures on thyroid disease co-existence detection can be found in Algorithm 4.

6.4 Experimental Design

Since Xception is the most accurate CNN model for the binary and multi-class classification tasks and is relatively efficient among the evaluated 11 models, it has been used as the backbone to implement the multi-channel architectures in this research.

More specifically, the SIDC architecture was evaluated through both CT and ultrasound images, while the DIDC and four-channel architectures were evaluated through CT scans. Additionally, the three architectures were incorporated to detect thyroid cancer for

Algorithm 4: Pseudo-code for implementing multi-channel CNN architectures

Inputs:
$X_L = \{X_1^a, X_2^a, .., X_n^a\}$; $X_R = \{X_1^b, X_2^b, .., X_n^b\}$
X is the input CT scan, n is total number of images, a denotes images from left-side, b denotes images from right-side
$y_L \in \{0, 1, 2, 3, 4, 5\}$; $y_R \in \{0, 1, 2, 3, 4, 5\}$
y is the corresponding class for image sets X_L and X_R
Divide into stratified $K - fold$ for training and testing
Initialisation:
Set $i = 0$; $K = 10$
for *k in K folds* **do**
  while $i \leq len(X)$ **do**
    Multi-channel CNN $\leftarrow (X_L, y_L; X_R, y_R)$; Input images and corresponding labels from both sides to the model simultaneously for training
    $p_i^a \leftarrow (X_L, y_L)$; $p_i^b \leftarrow (X_R, y_R)$; $p = p_i, p_{i+1}, .., p_{i+n}$; Predicted value indicating the class from 0 to 5 of the testing image
    if $p_i^a = 0$ *&* $p_i^b \neq 0$ *or* $p_i^a \neq 0$ *&* $p_i^b = 0$ **then**
      $Output \leftarrow p_i^b or p_i^a$; Final class of the patient would be the none 0 type
    else
      if $p_i^a = p_i^b$ **then**
        $Output \leftarrow p_i^a$; Final class of the patient would be P_i^a or p_i^b
      else
        if $p_i^a \neq p_i^b$ **then**
          $Output \leftarrow Fused(p_i^a, p_i^b)$
        end
      end
    end
    $i = i + 1$
  end
  Calculate correctly classified image in $k - fold$ using Eq. 5.4
end
Output: Averaged testing accuracy for K folds $\rightarrow \overline{Acc}$

the entire gland, while the four-channel model was further utilised for disease co-existence detection. This section describes the utilised datasets and presents the experimental parameter setting for each task.

6.4.1 Dataset Descriptions

In order to confirm the associations between kernel size and CNN performance, this research adopted both ultrasound and CT images for evaluation. More specifically, the SIDC architecture was evaluated with two image modalities. The patient-specific architectures (i.e., DIDC and four-channel) were evaluated with CT scans solely. The datasets are described in this section.

Ultrasound Dataset

The conventional single-channel CNN architectures were evaluated in Chapter 5. Thus, this section incorporates the same image sets for multi-channel CNN architectures for comparison. More specifically, the 448 DDTI images (i.e., 66 benign and 382 abnormal) and 917 CN ultrasound images (i.e., 717 benign and 200 malignant) were utilised for the SIDC architecture evaluation.

CT Dataset

For the evaluation of the SIDC architecture using the CT scans, a total number of $2,257$ segmented images from 578 patients were utilised, including $1,180$ left-side and $1,077$ right-side CT slices. This research aims to achieve thyroid disease co-existence diagnoses. The representative slices were selected rather than the entire volume of the CT. Therefore, 3 to 4 slices per lobe were limited in this work. Definitive diagnoses of each lobe based on FNAC or postoperative histopathology were used to label all the images. The images were excluded from this study if no histopathological results were available. The dominant class was assigned to the image based on the severity of the disease, in which cancer is regarded as the most severe type, followed by the adenoma

kind, multi-nodular goitre, cystic, thyroiditis, and normal type.

The DIDC and four-channel architectures have a rigorous prerequisite where this research needs to ensure the input sizes from two streams are the same and belong to one patient at a time. Therefore, the selection of the input images for DIDC and four-channel architectures was made manually. More specifically, Table 6.1 presents the distribution of the CT scans in the six classes for the DIDC and four-channel architectures, including 977 images from both sides.

Table 6.1: Distribution of CT scans in the six classes for DIDC and four-channel architectures.

Side	Class						Total
	Normal	Thyroiditis	Cystic	Goitre	Adenoma	Cancer	
Left	199	68	299	178	55	178	977
Right	217	72	308	179	47	154	977

6.4.2 Parameters Settings

This research considers the single-channel CNN architectures from Chapter 5 as the baseline for comparison. Therefore, the parameters were set as the same for consistency. In particular, the learning rate was set as 1×10^{-5}, and the batch size was set to 2 for the multi-channel models due to the expensive computational cost. The running environment was the same as in Chapter 5. The valuation metrics were also the same.

All the images were in the size of 224×224 in RGB channels, and the labels were annotated as $y \in \{0, 1\}$ for detecting benign or malignant class and $y \in \{0, 1, \ldots, 5\}$ for distinguishing among different types. More specifically, 0 to 5 denotes the "normal" to "cancer" class, which is the same as the multi-classification task. For the disease co-existence detection, the classes 6 to 15 indicate the combinations of the different diseases, which can be found in Figure 6.3.

The labelling process for disease co-existence detection was strict. In detail, if a patient has normal thyroid lobes for both sides of the gland, the patient is considered "normal". Similarly, if a patient has the same disease on both sides, the corresponding

images will be labelled with the dominating disease class. Moreover, if the patient has one side diagnosed as "normal" while the other side has a disease, the patient will be classified as the dominating disease class. For example, if the patient has a left-side CT image diagnosed as normal, and the right-side appears to have cancerous nodules, then the overall diagnosis would be "cancer". The different disease combinations were given to an individual patient based on the dominant class for each side of the CT image. Therefore, a total number of 16 classes were presented in this research. Besides the primary six classes, there are types include: thyroiditis with cystic nodule, thyroiditis with goitre, thyroiditis with adenoma, thyroiditis with cancer, cystic with goitre, cystic with adenoma, cystic with cancer, goitre with adenoma, goitre with cancer, and adenoma with cancer.

6.5 Results Analysis

This section examines the multi-channel architectures and interprets the ablation study results on four different kernel size combinations and gender disparity analysis to test its generalisation.

6.5.1 SIDC CNN Results

The SIDC architecture is dedicated to enhancing the diagnostic accuracy for thyroid disease by maintaining more features during the convolutional operations. This architecture incorporated the most commonly used kernel sizes (i.e., 3×3 and 7×7) for evaluation using ultrasound and CT image modalities. The single-channel and the dual-channel architectures were compared, and the results are demonstrated in Table 6.2.

For both image modalities, the SIDC architecture outperformed the single-channel models. More specifically, for the ultrasound images, the DDTI data reached a diagnostic accuracy of 0.987 through the CV technique with a variance of 0.001. A similar result was also presented for the CN ultrasound image set with an accuracy of 0.989 achieved.

With the CN CT scans, a slight increase was also found when applying the SIDC

Table 6.2: Single-channel and dual-channel performance comparison on ultrasound and CT images.

Dataset Modality	Architecture	Kernel Size	Accuracy
DDTI Ultrasound	Single	3	0.980 (±0.003)
	Single	7	0.984 (±0.018)
	SIDC	3, 7	0.987 (±0.001)
CN Ultrasound	Single	3	0.987 (±0.002)
	Single	7	0.988 (±0.003)
	SIDC	3, 7	0.989 (±0.003)
CN Left CT	Single	3	0.966 (±0.004)
	Single	7	0.972 (±0.007)
	SIDC	3, 7	0.975 (±0.008)
CN Right CT	Single	3	0.970 (±0.004)
	Single	7	0.974 (±0.013)
	SIDC	3, 7	0.975 (±0.005)

architecture, with both left and right lobes reaching an accuracy of 0.975. The pattern demonstrates that when the size of the convolutional kernel is altered, the performance of the model will be impacted. The multi-channel architectures demonstrated superior performance when combining the feature maps generated from 3×3 and 7×7 compared to the original embedded kernel size of 3 from the Xception model. When the kernel size changed to 7, a minor increase was presented for the accuracy in all datasets, and this might be accumulated when incorporating more images. Therefore, the proposed SIDC proves that enhanced diagnostic accuracy for thyroid disease can be attained.

6.5.2 DIDC CNN Results

Based on the SIDC architecture results, the DIDC architecture utilised the kernel combinations of 3×3 and 7×7 for implementation. The DIDC architecture is responsible for generating diagnoses for the entire thyroid gland following the patient-specific design. The DIDC architecture aims to produce a diagnostic decision for both sides of the thyroid gland for each patient at a time. The DIDC model firstly generated the binary classification decision for each side of the gland and merged the diagnoses made from each side into an overall decision, denoted from class 0 to 3, as explained in section 6.3.2. Table

6.3 presents the DIDC results where it generates promising performance in classifying normal (i.e., class 0) and malignant (i.e., class 3) patients with an F1 score of 0.91 and 0.97 correspondingly.

6.5.3 Four-Channel CNN Results

The four-channel architecture has been used for binary and multi-class classification tasks. The binary classification task indicates the overall thyroid gland status of the individual patient, whereas the multi-class classification task detects the disease co-existence phenomenon.

Binary-Class in Thyroid Gland Status Diagnosis

Table 6.3 presents the DIDC and four-channel performance on the binary task. More specifically, both architectures obtained competitive diagnostic results where the DIDC model reached an average accuracy of 0.95, and the four-channel model achieved an average accuracy of 0.94.

Comparing the results from four classes regarding the precision scores, the four-channel model has achieved a 1.00 accuracy rate for detecting "normal" patients, outperforming the DIDC model. When detecting abnormal patients, the DIDC architecture performed slightly better than the four-channel architecture. Those four classes were scored 0.88, 0.94, 0.91, and 0.97 for the four-channel architecture with the F1 metric.

Table 6.3: DIDC and four-channel architectures performance comparison on the binary classification task.

Architectures	Classes	Precision	Recall	F1
DIDC	0	0.97 (±0.091)	0.86 (±0.091)	0.91 (±0.091)
	1	0.99 (±0.088)	0.94 (±0.089)	0.96 (±0.089)
	2	0.97 (±0.089)	0.88 (±0.090)	0.92 (±0.090)
	3	0.96 (±0.006)	0.99 (±0.000)	0.97 (±0.002)
Four-channel	0	1.00 (±0.090)	0.78 (±0.110)	0.88 (±0.096)
	1	0.97 (±0.087)	0.91 (±0.089)	0.94 (±0.086)
	2	0.94 (±0.099)	0.89 (±0.089)	0.91 (±0.093)
	3	0.95 (±0.007)	0.99 (±0.001)	0.97 (±0.002)

Figure 6.4 shows the averaged accuracy scores for the 10-fold stratified CV with DIDC and four-channel architectures. The best performing results from each fold were averaged for the 10-folds (i.e., shown in black), where it can be argued that imbalanced data sets would result in some fluctuations from the CAD implementation. Besides these fluctuations, the proposed architecture is relatively stable and performs promising.

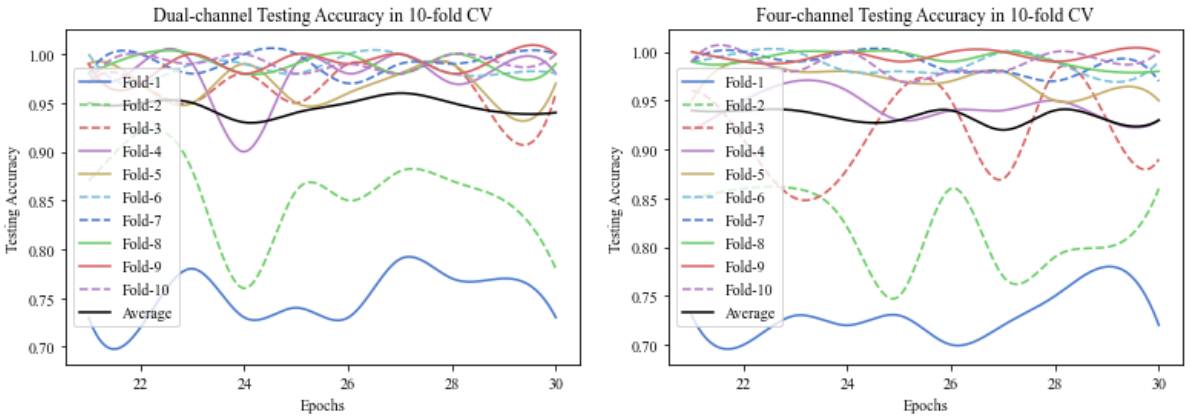


Figure 6.4: Averaged accuracy comparison of DIDC and four-channel architectures.

Multi-Class in Thyroid Disease Co-Existence Diagnosis

In order to further evaluate the different kernel size impacts on CNN performance, Table 6.4 presents the performance of the four-channel architectures in different kernel size combinations for the multi-class classification task tailored for thyroid disease co-existence detection.

When the 1×1 and 7×7 kernel size convolutions were applied, the highest accuracy among the four combinations was obtained, which was 0.909. The best F1 score was also achieved with this kernel setting. With the kernel combination of 3×3 and 5×5, the architecture achieved an accuracy of 0.906, which was slightly lower than that of the 1×1 and 7×7 combination. For the combination of the kernel sizes of the 3×3 and 7×7, the model obtained the lowest accuracy, precision, specificity, NPV, and F1 scores, which were 0.900, 0.907, 0.992, 0.992, and 0.903, respectively. Nevertheless, the highest recall score of 0.907 with a variance of 0.06 was achieved with 3×3 and 7×7 kernel combination.

Table 6.4: Multi-class classification performance of the four-channel architecture in detecting thyroid disease co-existence.

Kernel	Accuracy	PPV	Recall	Specificity	NPV	F1
1 & 5	$0.905 \pm .05$	$0.925 \pm .05$	$0.889 \pm .05$	$0.993 \pm .00$	$0.993 \pm .00$	$0.904 \pm .05$
1 & 7	$\mathbf{0.909 \pm .05}$	$\mathbf{0.944 \pm .06}$	$0.896 \pm .05$	$\mathbf{0.994 \pm .00}$	$\mathbf{0.994 \pm .00}$	$\mathbf{0.917 \pm .06}$
3 & 5	$0.906 \pm .05$	$0.918 \pm .05$	$0.894 \pm .05$	$0.993 \pm .01$	$0.993 \pm .00$	$0.904 \pm .06$
3 & 7	$0.900 \pm .06$	$0.907 \pm .06$	$\mathbf{0.907 \pm .06}$	$0.992 \pm .01$	$0.992 \pm .00$	$0.903 \pm .06$

Figure 6.5 displays the detailed F1 scores for the four kernel size combinations through the 10-fold stratified CV. The most stably performed architecture achieved the highest accuracy when the kernel sizes 1×1 and 7×7 combination were applied. More specifically, the 1×1 and 7×7 four-channel architecture reached mean F1 scores of higher than 0.9 for classifying 12 out of 16 classes, including normal, cystic, goitre, adenoma, cancer, thyroiditis with cystic, thyroiditis with adenoma, thyroiditis with cancer, cystic with goitre, cystic with adenoma, goitre with adenoma, and adenoma with cancer. In contrast, the lowest F1 score of 0.85 was obtained for classifying the cystic and cancer class. Besides, the lowest mean F1 scores of the four models were also presented for the thyroiditis with adenoma class. The combination of the kernel sizes 3×3 and 7×7 appeared to be the most fluctuating model.

6.5.4 Ablation Study

The ablation study was involved in this research to validate the enhanced performance of the multi-channel architectures by comparing them with single-channel architectures and gender disparity analysis to evaluate its generalisation.

Single-channel Comparison

The four-channel architectures were compared to the single-channel architectures (i.e., with the kernel size as of 3×3, 5×5, and 7×7), and Table 6.5 demonstrates their results. It should be noted that the kernel size of 1×1 was not evaluated in this case as it is usually used for dimension reduction [293]. In this regard, the disease co-existence circumstance

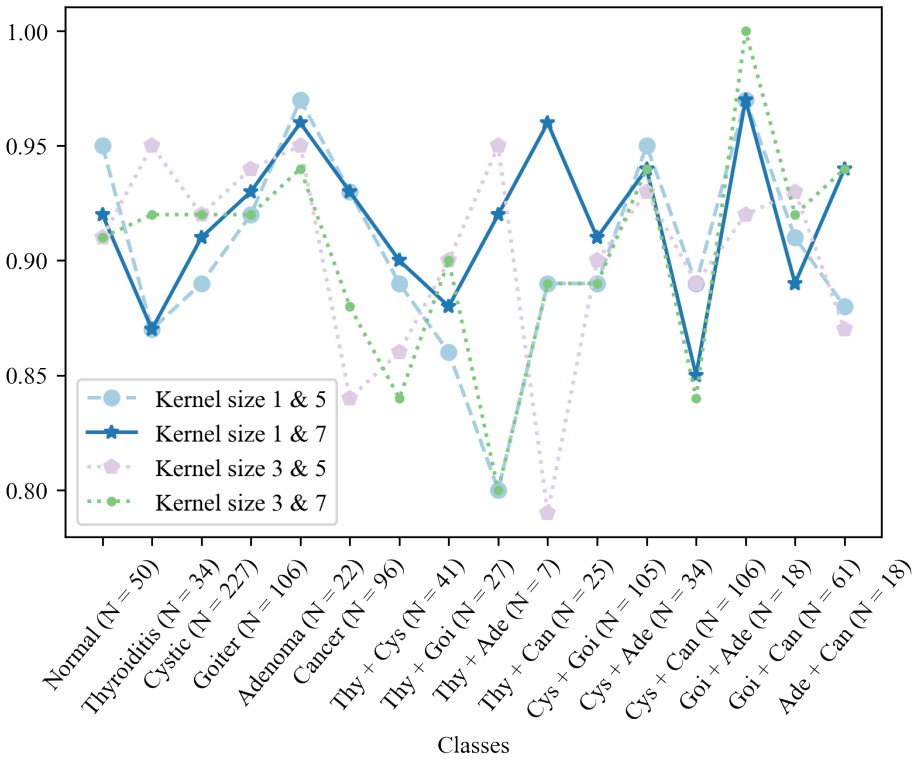


Figure 6.5: Mean F1 scores for the four-channel architecture on multi-classifying thyroid disease on 10-fold stratified CV.

was taken into consideration. Therefore, the classification results from the left and right lobes were fused.

Table 6.5: Multi-class classification performance of the single-channel architecture in detecting thyroid disease co-existence.

Kernel	Accuracy	PPV	Recall	Specificity	NPV	F1
3×3	$0.880 \pm .00$	$0.904 \pm .01$	$0.875 \pm .00$	$0.991 \pm .01$	$0.992 \pm .01$	$0.888 \pm .00$
5×5	$0.900 \pm .00$	$\mathbf{0.905 \pm .01}$	$0.894 \pm .00$	$0.992 \pm .01$	$\mathbf{0.993 \pm .00}$	$0.899 \pm .00$
7×7	$\mathbf{0.902 \pm .00}$	$0.892 \pm .01$	$\mathbf{0.909 \pm .00}$	$\mathbf{0.993 \pm .00}$	$\mathbf{0.993 \pm .01}$	$\mathbf{0.900 \pm .00}$

Figure 6.6 depicts the averaged CV performance comparison of the four-channel (i.e., kernel size of 1×1 and 7×7) and single-channel architecture (i.e., 7×7 kernel size). The four-channel architecture achieved increased accuracy, precision, specificity, NPV, and F1 scores compared to the single-channel architectures. The sensitivity of the four-channel architecture was slightly lower than the single-channel architecture, and its overall performance was superior and more stable.

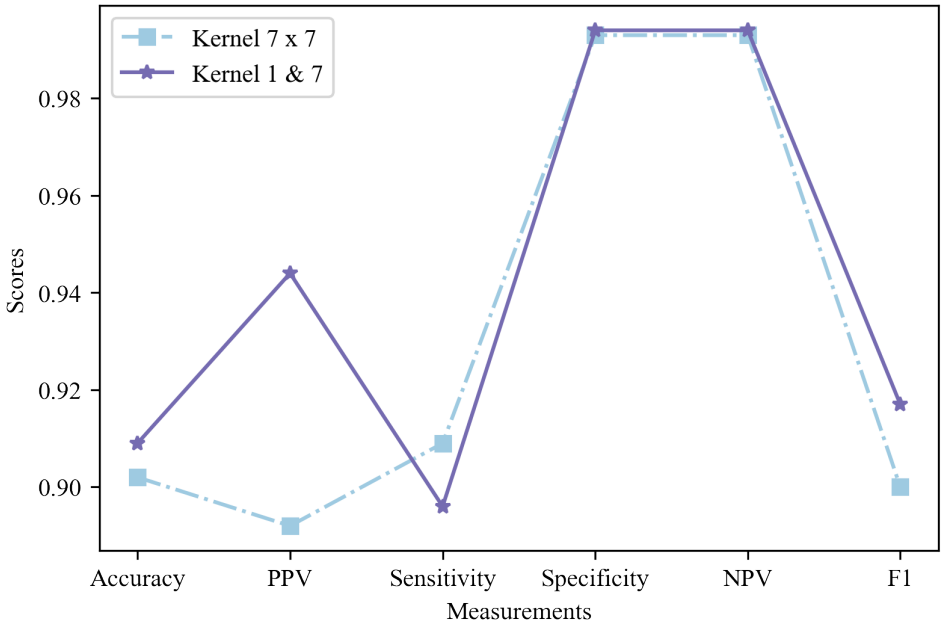


Figure 6.6: Averaged 10-fold CV performance metrics comparison of the four-channel and single-channel architectures.

Gender Disparity Analysis

Thyroid disease is more likely to be established in female groups. Therefore, in order to avoid developing a model prone to make accurate diagnoses for female patients, this section further evaluates the four-channel architecture of thyroid disease co-existence situation with different genders.

Table 6.6 exhibits the four-channel architecture (i.e., kernel size combination of 1×1 with 7×7) performance on female and male groups. The numbers of the input images for the female and male groups were 774 and 203, respectively. For both gender groups, the classification accuracy of the multi-channel CNN architecture is promising. For the female group, the proposed architecture reached 0.908, 0.931, 0.898, 0.994, 0.994, and 0.912 for accuracy, precision, recall, specificity, NPV, and F1, respectively. The male group had the corresponding scores of 0.901, 0.954, 0.9, 0.992, 0.992, and 0.913, respectively.

Figure 6.7 further displays the 10-fold CV averaged F1 scores for female and male groups in the 16 thyroid disease classes. Due to the data sample limitations, the classes "goitre with adenoma" and "adenoma with cancer" were absent for the male groups. More

Table 6.6: Gender disparity analysis.

Gender	Accuracy	PPV	Recall	Specificity	NPV	F1
Female	$0.908 \pm .01$	$0.931 \pm .00$	$0.898 \pm .01$	$0.994 \pm .01$	$0.994 \pm .00$	$0.912 \pm .00$
Male	$0.901 \pm .01$	$0.954 \pm .02$	0.900 ± 0.02	$0.992 \pm .00$	$0.992 \pm .00$	$0.913 \pm .01$

specifically, the proposed architecture generalises well to the different gender groups, with an F1 score larger than 0.85 for most classes. Although the input image scale for the male groups was much smaller, the model still achieved an outstanding performance. Therefore, the proposed architecture has the potential to be further extended to different diseases due to its excellent generalisation.

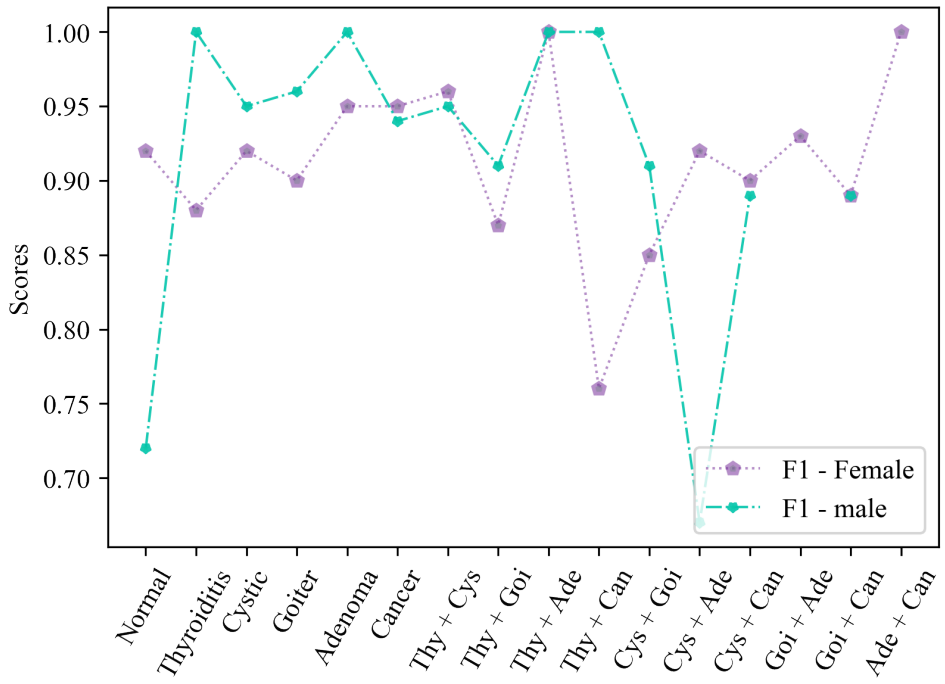


Figure 6.7: Mean F1 scores for the four-channel architecture on multi-classifying thyroid disease on the 10-fold stratified CV regarding gender disparity.

6.6 Discussion

The proposed multi-channel CNN architectures consist of three structures, whereas all were built by using Xception as the backbone. The three proposed architectures were evaluated using the real-world dataset, demonstrating superior performance to the

existing studies. Additionally, the choice of outputs (i.e., binary and multi-class) leads to highly interpretable results that generate potential adoptions in the clinical domain.

The proposed SIDC model obtained a diagnostic accuracy of 0.987 for the DDTI ultrasound images and 0.989 for the CN ultrasound image set. The results confirmed the strong association between kernel size and CNN performance. The accuracy rate of 0.975 was obtained for both left and right-side CT images through the SIDC architecture, all outperforming the single-channel architectures and existing works. The comparison details with existing works can be viewed in Table 6.7.

Table 6.7: Binary classification comparison with existing studies.

Modality	Methods	Image	Accuracy
Ultrasound	Zhu et al. [358]	DDTI	0.840
	Sundar et al. [290]	DDTI	0.940
	Raghavendra et al. [252]	DDTI	0.970
	Proposed SIDC	DDTI	**0.987**
CT	Li et al. [163]	832 scans	0.859
	Zhao et al. [345]	986 scans	0.874
	Lee et al. [158]	995 scans	0.904
	Zhao et al. [349]	1421 scans	0.957
	Proposed SIDC	2352 scans	**0.975**

The patient-specific DIDC and four-channel architectures are the novelties in this research and do not have any benchmark studies to compare. These two architectures also demonstrate outstanding classification results for the binary classification task. The DIDC model has achieved 0.95 diagnostic average accuracy, and the four-channel model has obtained a 0.94 average accuracy score. Besides, CT scans can reach comparable detection results compared to ultrasound images. This highlights the potential adoption of other medical image modalities for the implementation of CAD approaches.

Single-channel CNN models were typically used to distinguish between benign and malignant nodules [2], which are limited in considering the disease subtype co-existence situation. Accordingly, the four-channel architecture was further applied for a multi-classification task to address the disease co-existence detection issue. Among all the existing works in the literature on thyroid disease detection, the utilisation of CT scans

is considerably limited. Li et al. [163] deployed 832 thyroid CT slices to classify benign and malignant nodes and achieved an accuracy of 0.859. Zhao et al. [345] adopted 986 CT images and obtained an accuracy of 0.901. Masuda et al. [197] used CT images to identify thyroid cancer and achieved the area under the curve score of 0.86. Our four-channel architecture for multi-classifying thyroid disease is the first of its kind. It achieved the patient-specific diagnosis for interpreting the status of the entire gland, reaching an accuracy of 0.909 with kernel sizes of 1×1 and 7×7.

The comparative experiments were conducted among single and multi-channel architectures under different kernel size combinations. The 1×1 and 7×7 architecture generates the best performance with stability. The slight decrease of the single-channel compared to the multi-channel architectures in the sensitivity might be due to the increased model complexity, as the sensitivity will likely be affected and prone to over-fitting. Notably, the sensitivity score appeared relatively lower in all the experiments. The primary class affecting the CNN sensitivity score is "thyroiditis", in which the sensitivity score tends to be dragged down when this disease type exists. Such a result is expected and acceptable as thyroiditis can co-exist with neoplastic diseases and can confound the analysis. By considering labels given to each image based on the dominant class during the training stage, there might be cases where thyroiditis manifested on the image but is not labelled as that class, resulting in a lower sensitivity.

For the disease co-existence detection, the "goitre" class tends to exhibit the highest precision than the other classes for both single-channel and multi-channel architectures. The reason might be due to the multi-nodular goitre manifestation being distinct from the other types of nodules.

Despite the superior diagnostic performance of the proposed multi-channel CNN architectures, there are limitations related to the labelling process. In practice, each CT slice may have more than one diagnosis, and this has not been addressed by existing studies. In the proposed research, the dominating class of the image was given to each CT slice. It would be engaging in developing a new multi-class model to process multi-

labelled images with the k-hot encoded labels to provide a comprehensive diagnosis. In addition, although CT images can provide more details around the thyroid gland than ultrasonography, it is generally not the preferred modality to characterize thyroid diseases in the clinical setting. Clinicians may find it difficult to accept a thyroid CAD system based on CT images. Therefore, future studies may incorporate other medical imaging modalities to design patient-specific CAD systems.

6.7 Summary

This research proposes three adaptive multi-channel CNN architectures for different output choices, including binary classification denoting the status of the entire thyroid gland and multi-class classification detecting disease co-existence phenomenon.

The proposed SIDC, DIDC, and four-channel architectures were evaluated with real-life datasets. More specifically, the SIDC architecture exhibited increased diagnostic accuracy compared to single-channel architectures. The DIDC and four-channel architectures achieved patient-specific diagnosis and provided a comprehensive diagnosis for the entire thyroid gland. From the generated results, it was indicated that kernel sizes strongly impacted CNN performance. The intermediate size of the feature map was generated by combining one smaller and one larger kernel, which will eventually contribute to enhanced diagnostic accuracy. Furthermore, the three architectures demonstrated that the outputs were highly interpretative and easier to gain clinical trust.

The four-channel architecture was further evaluated with a multi-class classification task. Different kernel size combinations were analysed, and the model was tested with different gender groups, all demonstrating its promising generalisation.

This research bridges the literature gap regarding the detection of the thyroid disease co-existence situation. We envision more studies to be applied to rigorously examine the generalisation and performance of such multi-channel CNN architectures for diagnosing different diseases. With the increasing evidence of the feasibility of deep learning-based

approaches, clinicians may be more confident and comfortable working with artificial intelligence-based diagnostic tools to reduce their workloads and mitigate diagnostic bias or human false-positive rates.

Chapter 7

A Unified Model for Enhancements in Comprehensive Thyroid Disease Diagnostic Decision-Making

7.1 Introduction

Diagnostic decision-making in practice usually generates hypotheses that draw upon the clinician's expertise, associations, and experiences [79]. The key elements affecting the quality of diagnostic decisions are the professional's knowledge, cognitive capacities, and the patient records required for interventions [317]. Those factors contribute to diversified medical cultures. In clinical decision-making, it is generally required to incorporate one or more intuitive justifications for diagnosis to mitigate the degree of uncertainty aroused from incomplete mastery of knowledge [317]. Accordingly, involving more facets of diagnostic hypotheses will potentially enhance the quality of a decision being made.

Due to the increased sensitivity in the medical diagnostic procedures, over 50% of adults are determined to have thyroid nodules [3], while the clinical diagnostic process is rigorous yet complex. The overall diagnostic procedure accompanies the inevitable human false-positive and false-negative diagnostic rates. Although there is a large body

of existing works on incorporating medical imaging with deep learning in designing CAD systems for thyroid cancer detection [174, 53, 179], those models usually use unitary datasets, which is incompetent in generalising to different patient profiles. Therefore, this research proposes to use the ensemble modelling technique to utilise limited information in generating more objective, reliable, and comprehensive diagnostic decisions, intending to highlight the prospective impact of enhancements in thyroid cancer diagnosis.

The implementation of ensemble models is long-established with the emergence of deep learning techniques, and its performance tends to be more accurate and stable than the single models [352, 51]. The ensemble learning concept was introduced by Igelnik et al. [117] in 1999 to emphasise the generalisation mechanism. Since then, there has been an increasing number of ensemble-based designs in CAD applications [350, 304, 178, 125, 305]. Zheng et al. [350] proposed an ensemble model that combines 2D and 3D biomedical images for segmentation tasks. Velusamy and Ramasamy [304] evaluated three ensemble mechanisms to predict coronary artery disease, including average voting, majority voting, and weighted-average voting. Their work demonstrated that weighted-average voting qualifies as the best-performing mechanism with a 98.97% accuracy rate achieved. Loddo et al. [178] established the ensemble model by combining three CNNs through an averaging method to detect Alzheimer's disease through MRI, reaching a diagnostic accuracy of 99.29%. Many more similar studies were also proposed delicately aiming to detect various diseases more accurately [257, 254, 13, 209, 30, 1]. Figure 7.1 depicts that standard ensemble models that integrate results generated by multiple individual networks usually contribute to enhanced performance.

However, ascribing to the unitary input data source, those ensemble models have difficulty adapting to diverse patient profiles (e.g., with distinct demographics, ethnicity, and cultural habits). This effect, more or less, limits the comprehensiveness of diagnostic decision-making and its clinical applications. Therefore, the input variety and output results should be adequately considered to address the challenge.

Based on the conducted comprehensive literature review, only a few existing studies

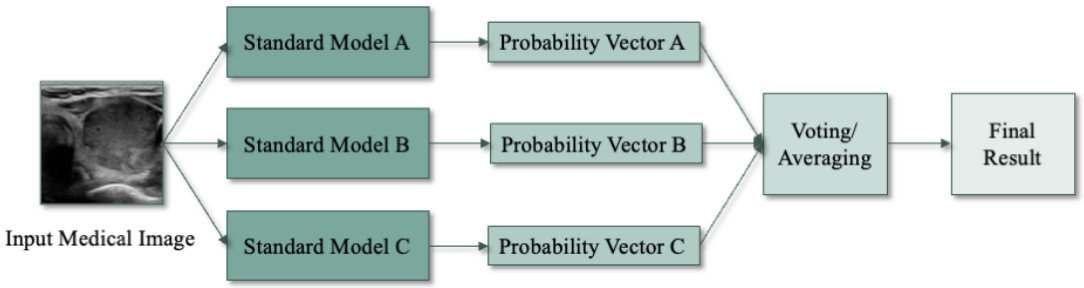


Figure 7.1: Classic ensemble modelling approach with unitary dataset used for training and testing in CAD design.

adopted the ensemble approach to enhance the detection accuracy for thyroid cancer diagnosis [53, 74]. Chen et al. [53] proposed a multi-view ensemble technique, which adopts major voting to make the final diagnostic decision by involving segmentation masks and original ultrasound images. Duc et al. [74] used the ensemble learning technique to aggregate the classification results performed from the FNAC slides in multiple classifiers. Although considered accurate, those ensemble modelling-based deep neural networks were usually trained and evaluated with unitary self-acquired datasets. This significantly impeded the models' ability to adapt to different institutions with diverse patient profiles. In this regard, the existing ensemble models cannot be utilised cross-institution, much less cross-nation. Accordingly, this research deployed data sources from two countries to build the unified model selection approach. It consists of a dynamic weighting mechanism and a weighted ensemble averaging model, strengthening its ability to be adaptive to diverse institutions, supporting reliable diagnoses, and further enlarging its clinical applications. In summary, this research makes contributions in the following four aspects.

- A unified model selection approach was proposed for utilising limited information in generating more objective, robust, reliable, and comprehensive diagnostic decision-making for thyroid cancer detection.

- The proposed model selection approach dynamically assigned weights to individual networks based on their pre-determined performance metrics, enhancing its adaptation to diverse data populations.

- A weighted ensemble averaging model was trained with cross-institutional data sources under diversified medical culture scenarios, tailored for precise thyroid cancer diagnosis.
- The proposed approach was dedicated to assisting clinicians with preliminary diagnostic decision-making and mitigating patients' financial and physical pressure from the clinical diagnosis. It can also be further applied to diagnose different diseases.

The contents of this chapter are in preparation for submission to the $26th$ ***Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD*** 2023***) in conference article*** 1***.***

7.2 Problem Formulation

The experimental results from the previous chapters show that ultrasound generates better diagnostic decisions than CT scans when designing CAD systems. This research then deployed the weighted ensemble averaging modelling technique to generate comprehensive and reliable thyroid cancer diagnostic decisions through ultrasound images. Data sources from two countries (i.e., Australian and Chinese datasets) were incorporated in training the model, and the external DDTI image set was used to further test the model.

Given the ultrasound image set from the Australian hospital (AU Dataset) $\mathcal{D}_A$ where $\mathcal{D}_A = \{X_a, y_a\}$, X_a indicates the encoded image matrix from the AU dataset, and $y_a \in \{0, 1\}$ as binary-class labels. Similarly, the ultrasound image set from the Chinese hospital (CN Dataset) is denoted as $\mathcal{D}_B = \{X_b, y_b\}$. The ultrasound image set from the DDTI repository (Open-access, OA Dataset) is represented as $\mathcal{D}_O = \{X_o, y_o\}$.

In order to reach consistency with the previous chapters, all the images from the three data sources were re-sized into 224×224. The proposed unified model selection approach selected the same baseline CNNs used earlier for evaluation (see Chapter 5).

7.3 The Proposed Unified Model Selection Approach

The proposed unified model selection approach consists of three main components (shown in Figure 7.2): *a*) the individual pre-trained CNN through cross-national data sources, *b*) the dynamic weight assignment mechanism for the ensemble averaging scheme, and *c*) the proposed weighted ensemble averaging model for more reliable and comprehensive diagnostic decision-making.

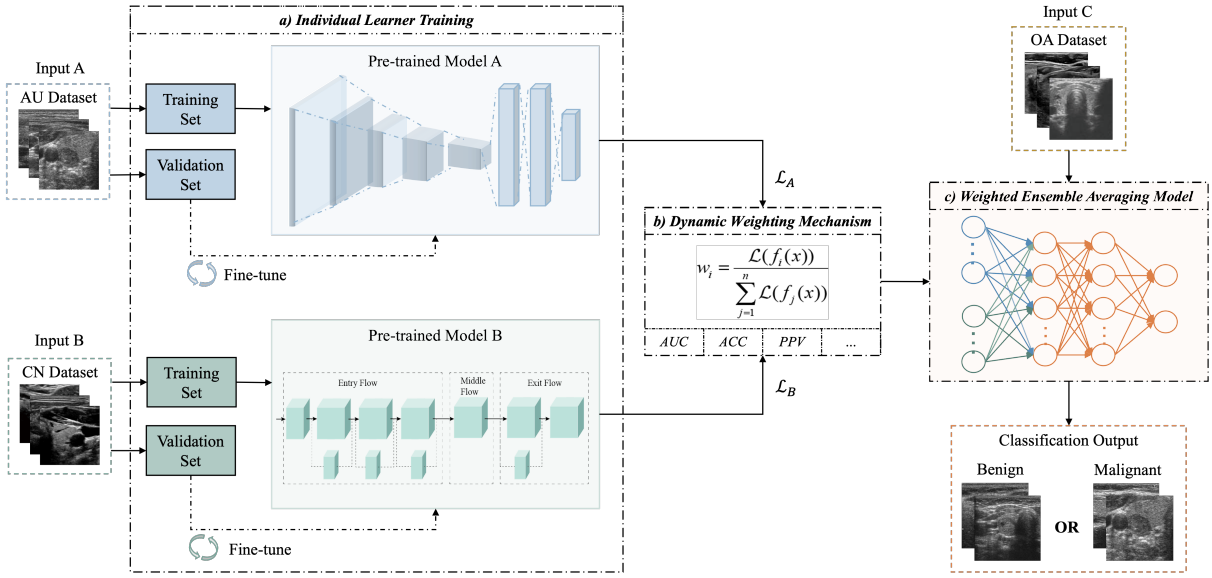


Figure 7.2: The proposed unified model selection approach.

The input cross-institutional datasets from two countries were split into training and validation sets separately. The individual CNN architectures were pre-trained using the cross-institutional training sets. The validation sets were used for fine-tuning the individual networks. The weights were assigned dynamically to the pre-trained individual networks based on their pre-determinate evaluation metrics, forming into the weighted ensemble averaging model. In addition, the proposed weighted ensemble averaging model was further evaluated with an external open-access dataset.

7.3.1 Individual Network Training

Given the No Free Lunch theorem (NFL) proposed by Wolpert and Macready [321], there is no single best-fitting CNN architecture for all datasets when they confront different image qualities, scales, and sizes. Therefore, this research selected the five most commonly utilised CNN architectures in the thyroid cancer detection task as candidate individual learners (i.e., individual networks) for the unified model selection approach [77, 312, 228, 313, 158]. The selected architectures include VGG11 [280], ResNet50 [105], DenseNet121 [115], InceptionV3 [293], and Xception [57]. Each baseline CNN architecture was independently evaluated and compared through pre-training and fine-tuning with the datasets acquired from two countries. Then, the best-performing candidate models were selected as individual learners from the two data sources. They were then assembled using a dynamic weighting mechanism to generate the unified ensemble model.

7.3.2 Dynamic Weighting Mechanism

There are three ensemble modelling schemes, including voting, averaging, and weighting [356]. In the existing studies, the voting and the averaging methods have been applied widely for computer vision tasks [350, 256]. The weighting method was introduced in 1998 by Jimenez [123] as aggregating outputs determined from several neural networks through a pre-determined weight. This mechanism emphasises that the weight would be higher when a neural network is more confident with its decisions being made. In other words, when an individual learner has high certainty in its predictive results, its weight should be higher than the others. This theory inspired our construction of the ensemble approach that the assignment of weights to the individual neural network should be dynamic and adaptive based on varied pre-determined performance criteria. Accordingly, when the individual learner performs better than the remaining models, its corresponding weights should be adjusted higher when assembling to the rest.

The weighted ensemble averaging model can be interpreted in Eq. 7.1, in which n is

the total number of models waiting to be assembled, w_i denotes the dynamically assigned weights to the individual learner based on its evaluation metrics, and f_i represents the performance results on the input image X.

$$f = \sum_{i=1}^{n} w_i f_i(X) \tag{7.1}$$

By considering its dynamic characteristic, w_i for the individual learner can be calculated as follows:

$$w_i = \frac{\mathcal{L}(f_i(x))}{\sum_{j=1}^{n} \mathcal{L}(f_j(x))} \tag{7.2}$$

More specifically, $\mathcal{L}$ denotes the metric value from the ith neural network. More precisely, the weights of the individual learners should be adjusted based on their predetermined performance criterion. In order to have a comprehensive analysis of the dynamic weighting distribution, this research has incorporated eight evaluation metrics, including area under a curve (AUC), accuracy (ACC), precision (i.e., positive predictive value, PPV), recall, specificity, F1, NPV, and false-positive rate (FPR, calculated with Eq. 7.3).

$$FPR = \frac{1}{K} \sum\nolimits_{i=1}^{K} \frac{FP_i}{FP_i + TN_i} \tag{7.3}$$

7.3.3 Weighted Ensemble Averaging Model

Zhou [356] once indicated that the individual learners should be accurate and diverse to propose an effective ensemble model. Deriving from this, this research pre-trained individual learners with cross-institutional data sources from two countries with distinct patient profiles under diverse medical cultural backgrounds. In this regard, the individual learners were constructed and selected to be accurate and diverse in their best-performing status. After assigning weights to the selected individual learners based on their predetermined performance metrics, the weighted ensemble averaging model was established.

In order to transfer the knowledge learnt from individual learners to the ensemble model, the transfer learning paradigm was incorporated in this case. Transfer learning has been well explored in existing studies for CAD implementations [352, 68]. It transfers knowledge learnt by a teacher network (i.e., individual learners) through a softened distribution of the final output to a student network (i.e., ensemble model). Given this protocol, the ensemble model can learn how the individual learners studied more effectively, given different data sources.

Intuitively, this research considered real-life clinical scenarios where cross-institutional datasets are of different qualities and scales to train and fine-tune the individual and ensemble models. More specifically, this research altered the size of the input image sources to mimic the reality in the clinical domain, including images with similar sample scales, one set over another in quantity, and the opposite. Extensive experiments were established with the adjusted image scales to evaluate the proposed model selection approach. Furthermore, this research evaluated the proposed ensemble models under different scenarios through the external DDTI dataset [243].

7.4 Experiments

With the involvement of two distinct data sources, the unified model selection approach can provide preliminary diagnostic decisions to support clinicians in detecting thyroid cancer. This section outlines the dataset descriptions and the weighted ensemble averaging model learning procedure.

7.4.1 Datasets Descriptions

With the ethics approval obtained, the CN and AU datasets were acquired. More specifically, 748 consecutive patients' electronic health records were utilised, including 578 patients from the Chinese hospital and 170 from the Australian hospital. The acquired records include their diagnostic reports and radiology images (i.e., ultrasound). Each of

the acquired ultrasound images was labelled based on its corresponding histopathological diagnosis. Otherwise, the images were removed if there were no pathological or cytological determinations.

Following the rigorous labelling process, 2,617 images were acquired from the Australian hospital (AU dataset), and 917 images were incorporated from the Chinese hospital (CN dataset). Notably, those images were acquired with different devices. The AU dataset utilised the Philips Medical Systems EPIQ and iU22 scanners, while the CN dataset involved a set of devices, including Philips EPIQ5, Hitachi, and S5. The AU dataset consists of a series of head and neck scans through the OHIF Viewer, while the CN images were embedded within the diagnostic reports. Therefore, more images were generated for the AU dataset than for the CN dataset.

The data augmentation techniques were applied to address the data imbalance issue in this case. Additionally, the data augmentation technique was also incorporated to alter the original image sets to mimic the diversified clinical scenarios for model evaluation. More specifically, the smaller scaled class (i.e., malignant) for both sets were rotated in 90°, 180°, 270° degrees. As a result, 4,339 AU and 1,517 CN images were generated. To further evaluate the generalisation of the proposed model selection approach, this research altered the size of the two datasets through sampling. The sampling process includes flipping for up-sampling and randomly selecting for down-sampling (i.e., three times up- and down-sampling), intimating the real-life scenarios that different institutions archive distinct image sample sizes. Table 7.1 presents the distribution of the raw, augmented, and altered datasets in each class.

Besides using the CN and AU datasets for pre-training the individual learner and constructing the weighted ensemble averaging model, this research used an open-access dataset (OA dataset) to evaluate the proposed ensemble model.

The OA dataset was obtained from the DDTI repository, which was proposed by Pedraza et al. [243]. The dataset consists of 448 ultrasound images from 400 patients. Those images were labelled following the TIRADS guideline, which classified images into

Table 7.1: Image class distribution in different scenarios.

Scenarios	Data Source	No. Benign	No. Malignant	Total
Raw	AU	2,043	574	2,617
	CN	717	200	917
Augmented	AU	2,043	**2,296**	**4,339**
	CN	717	**800**	**1,517**
Up-sampled	AU	2,043	2,296	4,339
	CN	**2,151**	**2,400**	**4,551**
Down-sampled	AU	**709**	**783**	**1,492**
	CN	717	800	1,517

seven risk stratification based on the number of suspicious features on images. To align with the existing reports in the literature [221, 290], this research has labelled the benign images with TIRADS ranking from 1 to 3, and the remaining ones ranked from 4 to 5 were labelled as malignant, which was the same labelling protocol from Chapter 5. Among the 448 images, 66 were labelled as benign and 382 as malignant. After augmenting the malignant class, 646 images were used for evaluations, including 264 benign and 382 malignant. Figure 7.3 illustrates some sample images and their corresponding descriptions from the three data sources with expert annotations and descriptions. Note, the OA dataset does not have expert region-of-interest annotations.

7.4.2 Learning Procedure

The overall learning procedure of the proposed unified model selection approach consists of two main phases, including pre-training the individual learners and transferring knowledge to the ensemble model through the dynamic weighting mechanism. Algorithm 5 interprets the pre-training and fine-tuning procedure of the individual learners. Algorithm 6 presents the detailed implementation of the weighted ensemble averaging model.

This research empirically evaluated the individual learners and the ensemble model with the Adam optimiser using the sparse categorical cross-entropy calculated in Eq. 7.4.

$$SCCE = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \tag{7.4}$$

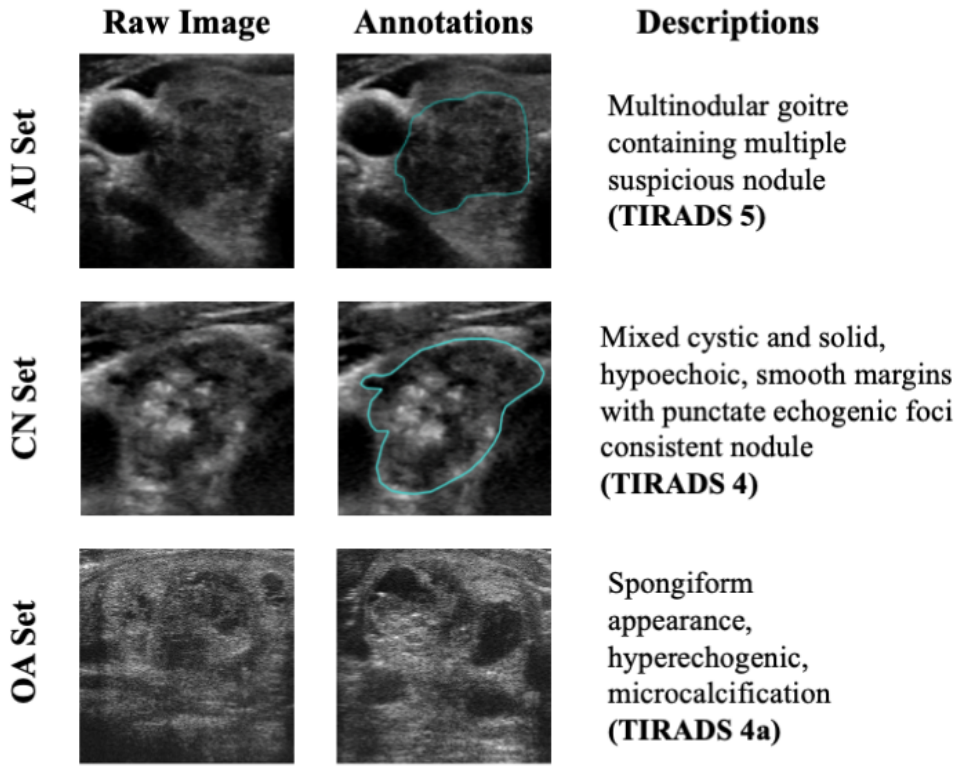


Figure 7.3: Sample images from the three data sources.

As this research involved independent data sources for training, validation, and testing, the 10-fold CV was not applied in this case. The training and validation split for the CN and AU dataset followed the ratio of $8:2$. During each training iteration, the learning rate was initially set to 1×10^{-3}, and it was gradually updated during the fine-tuning phase for each model. The batch size was set to 8 within 100 epochs, and the best-performing model was selected based on the validation accuracy.

7.5 Results

To evaluate the proposed unified model selection approach, this section interprets the results from three perspectives: individual learner's selection, weighted ensemble averaging model performance, and ablation study on distinctive weighting mechanism.

Algorithm 5: Individual learner selection

Input: $\mathcal{D}_A = \{X_a, y_a\}$, $\mathcal{D}_B = \{X_b, y_b\}$
Output: Performance metrics $\mathcal{L}_\mathcal{A}$ and $\mathcal{L}_\mathcal{B}$ of the best-performing individual learners
Split $\mathcal{D}_A$ and $\mathcal{D}_B$ into training and validation sets
for *augmented, up-sampled, and down-sampled* $\mathcal{D}_A$ *and* $\mathcal{D}_B$ **do**
  Pre-train individual networks $\mathcal{T}_\mathcal{A} = \{\mathcal{T}_A^1, \mathcal{T}_A^2, ..., \mathcal{T}_A^i\}$ and $\mathcal{T}_\mathcal{B} = \{\mathcal{T}_B^1, \mathcal{T}_B^2, ..., \mathcal{T}_B^j\}$ with training sets of $\mathcal{D}_A$ and $\mathcal{D}_B$
  Fine-tune individual networks $\mathcal{T}_\mathcal{A}$ and $\mathcal{T}_\mathcal{B}$ based on validation results
  while $\mathcal{T}_\mathcal{A}$ *and* $\mathcal{T}_\mathcal{B}$ *are with best performance* $\mathcal{L}_\mathcal{A} = \{\mathcal{L}_A^1, \mathcal{L}_A^2, ..., \mathcal{L}_A^8\}$ *and* $\mathcal{L}_\mathcal{B} = \{\mathcal{L}_B^1, \mathcal{L}_B^2, ..., \mathcal{L}_B^8\}$, *denoting metrics AUC, ACC, PPV, Recall, Specificity,* $F1$*, NPV, and FPR* **do**
    Record best-performing $\mathcal{L}_\mathcal{A}$ and $\mathcal{L}_\mathcal{B}$
  end
end
Store individual learners $\mathcal{T}_\mathcal{A}$ and $\mathcal{T}_\mathcal{B}$

Algorithm 6: Weighted ensemble averaging model evaluation

Input: $\mathcal{D}_O = \{X_o, y_o\}$, $\mathcal{T}_\mathcal{A}$, $\mathcal{T}_\mathcal{B}$, $\mathcal{L}_\mathcal{A}$, $\mathcal{L}_\mathcal{B}$
Output: Performance metrics $\mathcal{L}_\mathcal{E}$ of the weighted ensemble averaging model
Load individual learners $\mathcal{T}_\mathcal{A}$ and $\mathcal{T}_\mathcal{B}$
while $i = 1$ **do**
  Assign weights for $\mathcal{T}_\mathcal{A}$ and $\mathcal{T}_\mathcal{B}$ based on $\mathcal{L}_A^i$ and $\mathcal{L}_B^i$ through Eq. 7.2.
  Concatenate weighted $\mathcal{T}_\mathcal{A}$ and $\mathcal{T}_\mathcal{B}$ into a weighted ensemble averaging model $\mathcal{T}_\mathcal{E}$
  Input $\mathcal{D}_O$ into $\mathcal{T}_\mathcal{E}$ for evaluation with the eight metrics
  return $\mathcal{L}_\mathcal{E}$ of $\mathcal{T}_E$
  Store the ensemble model $\mathcal{T}_E$
  $i = i + 1$
end

7.5.1 Individual Learner Selection

Table 7.2 presents the performance of the candidate individual learners. The best-forming CNN architecture under the augmented situation for the AU dataset is the VGG11 model, reaching an AUC of 0.9811. While the best-performing model with the CN dataset is the Xception model, reaching an AUC score of 0.9156.

Table 7.3 illustrates the individual learner's performance under the up-sampled and down-sampled scenarios. Under all three scenarios, VGG11 demonstrated superior performance on the AU dataset, with 0.8255 accuracy rates for the up-sampled AU datasets.

Table 7.2: Individual learners' performance on AU and CN datasets (under the augmented scenario).

Model	Metrics							
	AUC	ACC	PPV	Recall	Spec	NPV	F1	FPR
AU Dataset								
VGG	**0.9811**	**0.9309**	**0.9498**	0.9163	**0.9469**	0.9116	**0.9327**	**0.0531**
Xception	0.9701	0.9217	0.9142	**0.9383**	0.9034	**0.9303**	0.9261	0.0966
ResNet	0.8937	0.8203	0.8565	0.7885	0.8551	0.7867	0.8211	0.1449
Inception	0.9002	0.8433	0.8219	0.8943	0.7874	0.8717	0.8565	0.2126
DenseNet	0.8969	0.8272	0.7774	**0.9383**	0.7053	0.9125	0.8503	0.2947
CN Dataset								
VGG	0.8916	0.8158	0.8690	0.8111	0.8226	0.7500	0.8391	0.1774
Xception	**0.9156**	**0.8618**	**0.9481**	0.8111	**0.9355**	0.7733	**0.8743**	**0.0645**
ResNet	0.7348	0.7895	0.8222	0.8222	0.7419	0.7419	0.8222	0.2581
Inception	0.7758	0.7829	0.8065	0.8333	0.7097	0.7458	0.8197	0.2903
DenseNet	0.7958	0.7961	0.7980	**0.8778**	0.6774	**0.7925**	0.8360	0.3226

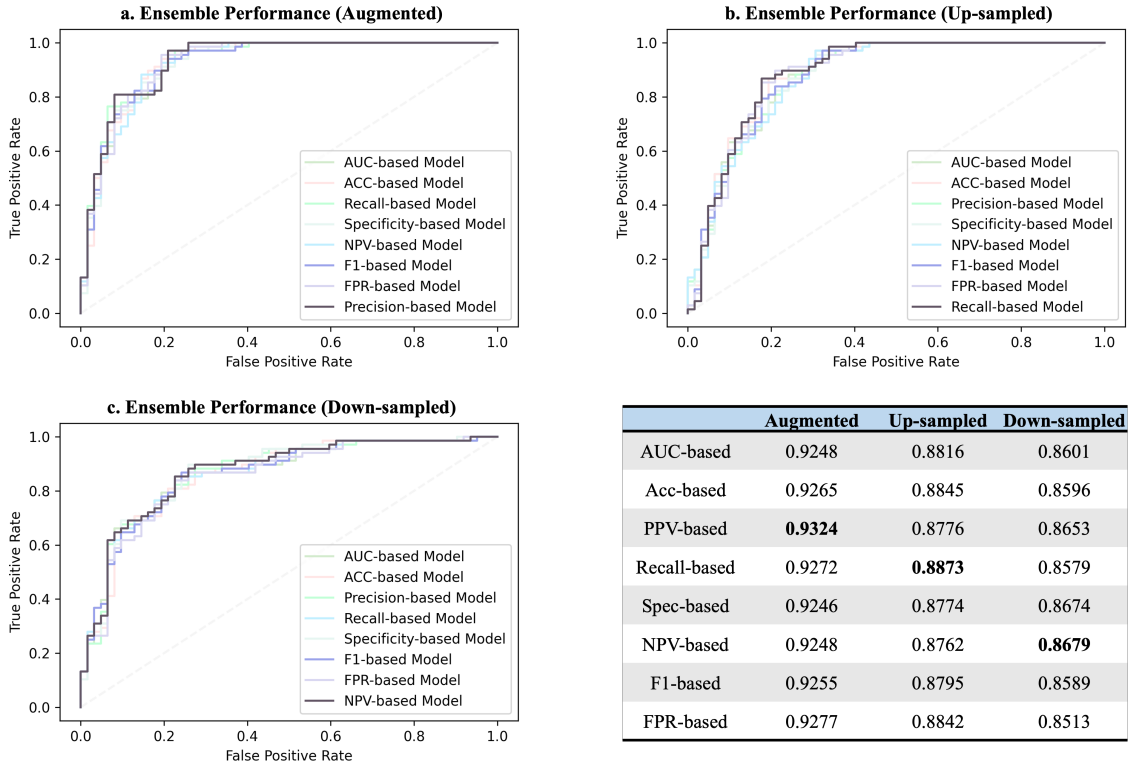
Similarly, Xception outperformed all other models on the CN dataset, reaching an accuracy of 0.8901 in the up-sampled dataset. Based on the performance metrics from the three scenarios, the weights of the individual learners were assigned for constructing the weighted ensemble averaging model.

Table 7.3: Individual learners' performance on altered dataset (under the up-sampled and down-sampled scenarios).

Model	Metrics							
	AUC	ACC	PPV	Recall	Spec	NPV	F1	FPR
Down-sampled AU Dataset								
VGG	**0.8603**	**0.8255**	0.8404	**0.8778**	0.7458	0.8000	**0.8587**	0.2542
Xception	0.8458	0.8054	**0.8506**	0.8222	**0.7797**	0.7419	0.8362	**0.2203**
ResNet	0.7242	0.7651	0.7570	0.9000	0.5593	0.7857	0.8223	0.4407
Inception	0.7665	0.7651	0.7570	0.9000	0.5593	0.7857	0.8223	0.4407
DenseNet	0.7470	0.7987	0.7830	0.9222	0.6102	**0.8372**	0.8469	0.3898
Up-sampled CN Dataset								
VGG	0.5346	0.5516	0.5524	0.5957	0.5067	0.5507	0.5732	0.4933
Xception	**0.9538**	**0.8901**	**0.9327**	**0.8435**	**0.9378**	**0.8543**	**0.8858**	**0.0622**
ResNet	0.6883	0.6659	0.7143	0.5652	0.7689	0.6337	0.6311	0.2311
Inception	0.6870	0.6681	0.8160	0.4435	0.8978	0.6121	0.5746	0.1022
DenseNet	0.8761	0.7934	0.8238	0.7522	0.8356	0.7673	0.7864	0.1644

7.5.2 Ensemble Model Performance

After assembling the individual learners selected from each pre-training iteration, the weighted ensemble averaging model was evaluated with the OA dataset to demonstrate its generalisation under the pre-defined diversified situations. Figure 7.4 illustrates the AUC values for the weighted ensemble averaging model in the three distinctive scenarios.



	Augmented	Up-sampled	Down-sampled
AUC-based	0.9248	0.8816	0.8601
Acc-based	0.9265	0.8845	0.8596
PPV-based	**0.9324**	0.8776	0.8653
Recall-based	0.9272	**0.8873**	0.8579
Spec-based	0.9246	0.8774	0.8674
NPV-based	0.9248	0.8762	**0.8679**
F1-based	0.9255	0.8795	0.8589
FPR-based	0.9277	0.8842	0.8513

Figure 7.4: AUC curve for the weighted ensemble averaging model on the open-access dataset under the different scenarios.

More specifically, the AUC score for the weighted ensemble averaging model under the augmented situation is 0.9248, and for the up-sampled and down-sampled situations is 0.8816 and 0.8610, respectively. Under all three cases, the ensemble model demonstrated satisfying adaption with all the AUC values reaching higher than 86% and accuracy values higher than 85%. Among the three situations, the augmented scenario demonstrated the highest AUC values greater than 92%. After altering the original image set, a slight decrease in AUC was found in both up-sampled and down-sampled datasets. This again supports the statement by Zhou [356] that ensemble models cannot simultaneously ensure

their accuracy rates and diversity. We might need to sacrifice accuracy when trying to enhance the model's diversity.

7.5.3 Ablation Study

In order to evaluate the best strategy for assigning weights to the ensemble model, the ablation study involved analysing the weighted models constructed based on the eight different pre-determined performance metrics. Figure 7.4 also exhibits the AUC values under different weight assignment strategies.

For the augmented situation, the best-performing strategy was when applying the precision-based weighting scheme, with an AUC of 93.24% achieved. Under the up-sampled situation, the best AUC was obtained through the recall-based weighting strategy, with a value of 88.73%. As far as the down-sampled case is concerned, the best-performing mechanism was the NPV-based weighting strategy, with an AUC of 86.79%.

Heuristically, the ensemble model demonstrated comparable performance when adopting dynamic weight assignment strategies. Moreover, the ability to be adapted to diversified input volume streams also showcases the ensemble model's promising performance. This further highlights its generalisation as it always finds a balance between the better and the worse performing models, making it unified and adaptive to different data sources.

7.6 Discussion

To yield a general adaptive model to support clinicians in detecting thyroid cancer requires the diversity of the input sources. Different institutions have distinct protocols for archiving patient records and managing electronic health systems, which is even conspicuous across nations with diversified medical cultures. Accordingly, this aggravates the challenge of CAD's applications in the clinical setting. Moreover, nations with developed characteristics tend to have confidence in their established guidelines, systems, and architectures, thus, adopting diagnostic decisions generated by their designed CAD systems.

On the other hand, developing countries usually follow protocols and adopt CAD systems created by developed nations. In this regard, such a protocol or a CAD system might not be suitable for those developing countries considering the distinct patient characteristics and the quality of their records (i.e., scale and feature). The proposed unified model selection approach sought to enrich medical culture by enhancing comprehensive diagnostic decision-making from diversified sources.

During the individual learner's selection phase, our results supported the NFL theorem [321] that there is no best-performing architecture in all types of data sources. This also emphasises the use of ensemble models for comprehensive diagnostic decision-making. Sometimes if the diversity of a model has to be increased, the accuracy should be sacrificed [356]. With the proposed weighted ensemble averaging model, this research sought to mitigate such a risk and provide precise and effective diagnostic decisions for patients from diverse cultural backgrounds with distinct demographic features. Since this work is the first of its kind, which incorporates cross-national data sources for the ensemble CAD model construction in the thyroid cancer diagnosis domain, there is no existing benchmark to be compared. Therefore, this research deployed three scenarios to evaluate the proposed weighted ensemble averaging model. With the augmented situation, the AU and CN datasets were in different quantities as $4,339$ and $1,157$, respectively. This was when the ensemble model produced the highest AUC values on the OA dataset of 0.9248 ± 0.0076 for the eight metrics-based weighting strategies. However, after altering the original datasets with up-sampling and down-sampling schemes to match the two datasets in quantity, the AUC decreased slightly to 0.9 ± 0.02 on the OA dataset. The results indicated that rotation as an image augmentation technique performed better than flipping. Accordingly, during the implementation stage of the ensemble model, it is recommended to avoid adjusting the original sample size to match the other source, as this will inevitably influence the model's performance.

Concerning the weighting strategy selection, the eight pre-determined performance metrics obtained relatively comparable performance, which was not surprising. Based on

the dynamic weighting scheme during the knowledge transferring phase, the weights of the individual learners determined the confidence and certainties of their outputs [123]. According to the performance metrics from the individual learners, their results were competitive with each other, leading to relatively equally distributed performance after assembling. The dynamic weighing strategy ensured the self-adaptation of the proposed approach as the weights were assigned each time during pre-training iteration rather than a fixed determination. In this regard, the ensemble model can be adaptive to different data sources from the beginning of the individual model selection stage until the weighted ensemble model construction. This highlights the diversity of the input data sources and supports accurate diagnostic decision-making under imbalanced medical culture.

In summary, with the proposed unified model selection approach, this research highlights using limited information to improve the reliability and comprehensiveness of diagnostic decision-making. Future research plans suggest including multiple data sources from diversified medical cultures to build the ensemble model for generating more sophisticated, reliable, and robust diagnostic decisions.

7.7 Summary

In summary, this research study proposes a unified model selection approach for constructing a weighted ensemble averaging model that can be adaptive to diversified data sources.

The proposed model selection approach consists of (1) a self-directed individual learner selection mechanism, (2) a dynamic weighting scheme based on eight pre-determined performance metrics, and (3) a general adaptive ensemble model which unifies CNN trained from cross-national datasets.

The proposed model was assessed to detect thyroid cancer through ultrasound image sets from three data sources. The individual learner selection involves using the five most commonly seen CNN architectures. The proposed weighted ensemble averaging model

was evaluated under three distinct clinical scenarios through the DDTI dataset. Extensive experiments indicated the promising performance of the proposed ensemble model. The proposed model utilised limited information to generate more comprehensive, robust, and reliable diagnostic decisions for patients with thyroid nodules.

Furthermore, this research provides recommendations for ensemble-based CAD implementations across nations to enrich medical culture. In future research, applying the ensemble model to other imaging modalities is suggested and incorporating more data sources for assembling to enhance the diversity of the model.

Chapter 8

Conclusion and Future Plan

8.1 Conclusion

Thyroid disease is a highly pervasive endocrine disease, and thyroid cancer is rising at the fastest rate among all malignancies [303]. Over 50% of adults have thyroid nodules, while such neoplastic thyroid disease is typically painless and undetectable by patients themselves [224, 3]. Due to the advances in sensitive medical imaging and the increased rate of regular health examination mechanisms, thyroid cancer instances are in a continuously rising pattern [303, 283].

Although the disease instances are increasing yearly, the epidemiology cause is still unknown. A substantial number of studies adopted qualitative techniques and statistical analysis to identify the risk factors associated with the disease. The limitation is that they usually investigate one factor at a time and have considerably ignored the associations among diverse risk attributes.

Neoplastic thyroid disease diagnosis in the clinical setting is costly, time-consuming, and stressful for patients. Each diagnostic examination is correlated with varying degrees of uncertainties in human false-positive and false-negative rates. This may contribute to the increased rates of unnecessary FNAC, excisional biopsy, or operation.

CAD designs for thyroid disease detection usually rely on a unitary dataset for evalua-

tion, which restricts the comprehensiveness in relevant diagnostic decision-making, while the implementation of such types of models cannot be generalised to different data sources with distinct patient profiles. Those challenges aggravate the limited clinical applications of machine learning-based approaches.

Accordingly, this thesis proposed three research questions extracted from the comprehensive literature review analysis, addressing the literature gaps and clinical challenges from three perspectives. The findings and novelties of this thesis can be summarised as follows:

- *A comprehensive literature review framework was proposed to identify research gaps and can be adapted to different research disciplines (Chapter 2).*

  The proposed comprehensive literature review framework offers a structured organisation of the existing studies to examine the historical and recent state-of-the-art machine learning approaches around the pathogenesis, diagnosis, and prognosis of thyroid cancer. Current challenges faced by clinicians and computer science approaches were also covered in this study. This framework has the potential to be adapted to multi-disciplines for conducting literature analysis.

- *The TM-DM framework was proposed to identify and evaluate risk factors correlated with thyroid disease (To achieve objective* 1*, Chapter 4).*

  Existing studies generally investigate a single factor at a time to identify its association with thyroid disease development through qualitative or statistical techniques [43, 103]. Such techniques are inefficient and neglect the correlations among risk factors. The TM-DM framework proposed in this thesis incorporates text and data mining procedures to evaluate risk factors obtained from two data sources in identifying thyroid disorders and cancer causes. The proposed framework allows extracting common and exception rules simultaneously (e.g., comorbidity and medical history) from raw health records through a CPIR measurement. Through ARM, feature selection, and classification tasks, a set of risk factors have been confirmed related to

thyroid disease development, including age, gender, thyroxine intake status, history of thyroid disease, history of thyroid surgery, hypertension, and thyroid function measurements (i.e., TSH, T3, FTI, FT3, FT4).

- *A CNN-BM framework was introduced to mitigate human false-positive and false-negative diagnostic rates (To achieve objective* 2*, Chapter 5).*

  This research is the first of its kind which incorporates two pre-operative medical image modalities to develop CNN-based CAD systems through binary and multi-class classification tasks. The pre-operative image modalities achieved automatic diagnosis and elevated clinical diagnostic accuracy compared to existing works [220, 9, 300, 227, 242]. The multi-class classification also stimulates reaching an expert-level diagnosis for CAD models. The implementation of 11 CNN architectures was conducted and the results from ultrasound and CT images were interpreted and compared. Both image modalities reached promising diagnostic accuracy, indicating the potential application of other image modalities in CAD designs.

- *The MC-TDD framework was proposed to elevate CAD accuracy and reach patient-specific design in thyroid disease detection (To achieve objective* 2*, Chapter 6).*

  The MC-TDD framework introduces three novel multi-channel CNN architectures tailored for precise thyroid disease detection, including SIDC, DIDC, and four-channel structures. The SIDC is dedicated to increasing diagnostic accuracy by combining feature maps generated from different kernel size convolutional operations, reaching the state-of-the-art diagnostic accuracy compared to existing works [4, 284]. The DIDC architecture is the first kind of CAD system which reaches the patient-specific design, allowing a diagnostic decision to be made for one patient at a time, diagnosing the status of the overall gland. The four-channel architecture combines the SIDC and DIDC benefits and is further evaluated through a multi-class classification task. This research is the first study which addressed the thyroid disease co-existence phenomenon, accomplished through the four-channel architecture,

demonstrating promising accuracy.

- *A unified model selection approach was proposed to use limited information in generating comprehensive and reliable diagnostic decisions for patients with thyroid disease (To achieve objective* 3*, Chapter 7).*

  The existing ensemble models were designed by incorporating classification results produced by diverse models, whereas those models have limitations in generalising to different data sources. Accordingly, this research introduced the unified model selection approach, which was the first of its kind that utilised image sets acquired from two countries to pre-train the individual networks. The best-performing pre-trained individual learners were assembled through a dynamic weighting mechanism into a weighted ensemble averaging model. The ensemble model was further evaluated with an external dataset. Extensive experiments demonstrated the promising generalisation of the approach under different clinical scenarios. This process learned decisions made from cross-institutions cross-nationally, enhanced the diversity of the ensemble model and made it adaptive to different data sources with distinct patient profiles.

Overall, this thesis extends our understanding of the current under-researched area of thyroid disease pathogenesis. Ultimately, this thesis aims to help mitigate patients' financial and mental pressure from clinical diagnosis while assisting clinicians in diagnostic decision-making.

8.2 Future Plan

Besides understanding thyroid cancer epidemiology and assisting clinicians with preliminary diagnostic decision-making, the treatments and prognosis of the disease are equally important. However, due to research scope limitations, the treatment recommendation systems and the prediction of disease prognosis are the two primary focuses in future research planning.

8.2.1 Treatment Recommendation Systems

Thyroid cancer risk stratification depends on its subtypes. More specifically, papillary carcinoma is the least severe kind, and anaplastic is the most severe kind, which is defined based on their mortality rates [204]. Different subtypes expect to undergo varied treatment protocols. Treatment protocols need to be developed rely on the individual patient's health conditions, such as age, weight, BMI, pregnancy status, medical history, and medication doses for other diseases. Sometimes, external factors like seasonal temperature change, financial status, and patient preference also play significant roles in establishing treatment plans [272]. Moreover, over-treatments are substantially occurring with thyroid disease [12, 157, 121, 269], and this usually leads to irreversible damage to the salivary gland or even the endocrinology system. To avoid such post-treatment adverse effects, customised treatment protocols must be established to achieve precision medicine and treatment for the target patient.

The issue around the existing machine learning-based treatment recommendation systems is that those models were designed as a general platform rather than explicitly built for thyroid disease. More importantly, the evaluations of those treatment recommendation systems were significantly ignored by the scientific community. We believe that there is a significant demand for treatment recommendation systems specifically built for thyroid disease in future works. In order to assist clinicians in establishing treatment recommendation protocols, more comprehensive factors besides the ones listed in the guidelines should be taken into consideration, such as tumour characteristics, medical history, and comorbidity.

The evaluations of the designed customised treatment recommendation systems must be arisen to enhance the potential adoption in the clinical domain. This can be achieved by clinicians' involvement in accelerating the patient-centric design and adoption of such systems. Figure 8.1 depicts a sampled FCM rule-based customised treatment recommendation system which considers individual patient profiles independently and incorporates

medical experts' opinions in the decision-making process. The generated recommendations include the choice of surgery, surgery type (i.e., total, partial, or completion thyroidectomy), daily thyroxine medication intake level, radiation therapy choice, or even Iodine-131 dosage. The system is potentially to be delivered through a digital platform, which supports clinicians in a rapid decision-making iteration, meanwhile allowing patients and experts access to a self-driven decision support system for helping to improve the prognosis.

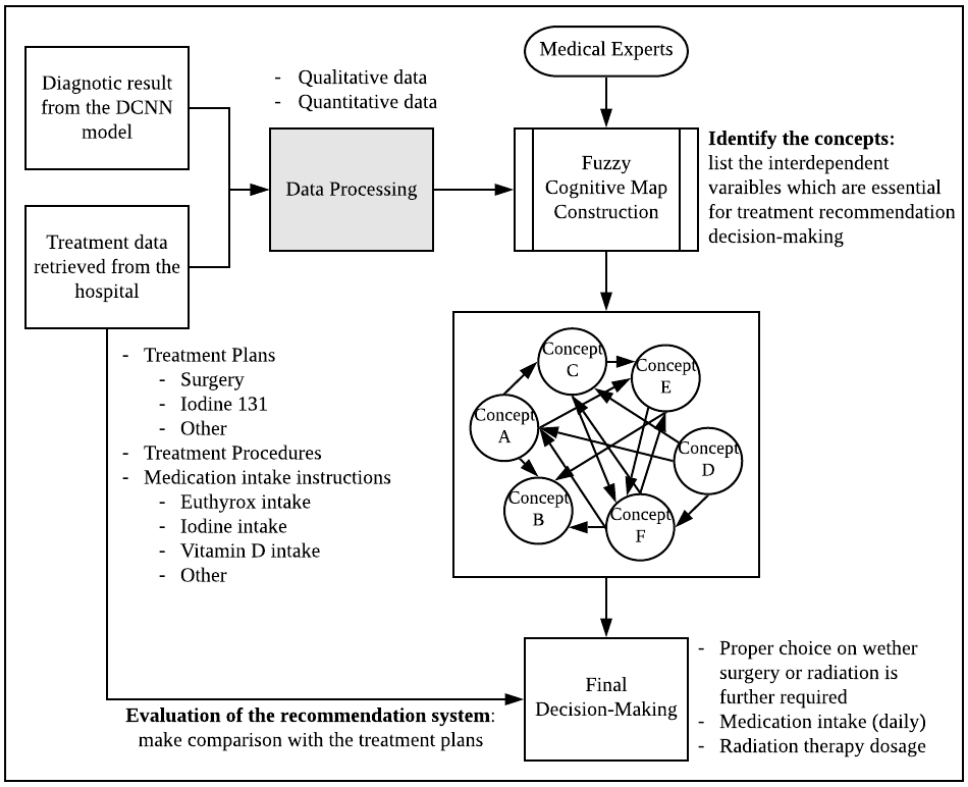


Figure 8.1: FCM-based customised treatment recommendation system.

8.2.2 Survival, Death, and Recurrence Prediction

The prediction of cancer survival, death, and recurrence rates is another topic on hit, and examples can be found in [75, 93, 270, 311, 331, 240, 265]. Extensive studies utilised patients' demographic features, tumour characteristics, and lymph metastasis factors to make predictions, yet ignoring the "time" factor.

In order to develop algorithms that can predict the prognostic status, time is particularly imperative as clinicians would potentially prepare for the upcoming health changes appearing in the patient. Therefore, the future research plan will integrate RNN, in particular, Long Short-Term Memory (LSTM), and ANN that considers time to make related predictions so that patients can better understand their health status. Accordingly, clinicians can be well prepared for the health condition changes appearing in the target patient regarding the subsequent period intervals to guide patients to achieve an optimal health record for an ideal prognosis.

$\mathbb{E}.\mathbb{N}.\mathbb{D}$

References

[1] Heba Abdel-Nabi, Arafat Awajan, and Mostafa Ali. A novel ensemble strategy with enhanced cross attention encoder-decoder framework for tumor segmentation in whole slide images. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 262–269, 2022. doi: 10.1109/ICICS55353.2022.9811163.

[2] Fatemeh Abdolali, Jeevesh Kapur, Jacob L. Jaremko, Michelle Noga, Abhilash R. Hareendranathan, and Kumaradevan Punithakumar. Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. *Computers in Biology and Medicine*, 122:103871, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.103871.

[3] U. Rajendra Acharya, G. Swapna, S. Vinitha Sree, Filippo Molinari, Savita Gupta, Ricardo H. Bardales, Agnieszka Witkowska, and Jasjit S. Suri. A review on ultrasound-based thyroid cancer tissue characterization and automated classification. *Technology in Cancer Research & Treatment*, 13(4):289–301, 2014. doi: 10.7785/tcrt.2012.500381. PMID: 24206204.

[4] Oluwadare Adepeju Adebisi, John Adedapo Ojo, and Temitope Olugbenga Bello. Computer-aided diagnosis system for classification of abnormalities in thyroid nodules ultrasound images using deep learning. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22(3):60–66, 2020.

[5] Oluwadare Adepeju Adebisi, John Adedapo Ojo, and Temitope Olugbenga Bello. Computer aided diagnosis system for classification of abnormalities in thyroid nodules ultra-

sound images using deep learning. *Journal of Computer Engineering*, 22:60–66, 2020. doi: 10.9790/0661-2203016066.

[6] Laura Agate, L Lorusso, and Rossella Elisei. New and old knowledge on differentiated thyroid cancer epidemiology and risk factors. *Journal of endocrinological investigation*, 35 (6 Suppl):3–9, 2012.

[7] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Arnab Kumar Laha, editor, *Proc. 20th int. conf. very large data bases*, volume 1215, pages 487–499. VLDB, 1994.

[8] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, page 207–216, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897915925. doi: 10.1145/170035.170072.

[9] OA Ajilisa, VP Jagathyraj, and MK Sabu. Computer-aided diagnosis of thyroid nodule from ultrasound images using transfer learning from deep convolutional neural network models. In *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, pages 237–241. IEEE, 2020.

[10] LA Akslen, S Nilssen, and G Kvåle. Reproductive factors and risk of thyroid cancer. a prospective study of 63,090 women from norway. *British journal of cancer*, 65(5):772–774, 1992.

[11] N. Aldoj, S. Lukas, M. Dewey, and T. Penzkofer. Semi-automatic classification of prostate cancer on multi-parametric mr imaging using a multi-channel 3d convolutional neural network. *European radiology*, 30(2):1243–1253, 2020.

[12] Erik K. Alexander and P. Reed Larsen. High dose 131i therapy for the treatment of hyperthyroidism caused by graves' disease. *The Journal of Clinical Endocrinology & Metabolism*, 87(3):1073–1077, 03 2002. ISSN 0021-972X. doi: 10.1210/jcem.87.3.8333.

[13] Redha Ali, Russell C. Hardie, Barath Narayanan Narayanan, and Supun De Silva. Deep learning ensemble methods for skin lesion analysis towards melanoma detection. In *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, pages 311–316, 2019. doi: 10.1109/NAECON46414.2019.9058245.

[14] Soo-Youn An, So Young Kim, Dong Jun Oh, Chanyang Min, Songyoung Sim, and Hyo Geun Choi. Obesity is positively related and tobacco smoking and alcohol consumption are negatively related to an increased risk of thyroid cancer. *Scientific reports*, 10(1): 1–9, 2020. doi: https://doi.org/10.1038/s41598-020-76357-y.

[15] Vatsala Anand and Deepika Koundal. Computer-assisted diagnosis of thyroid cancer using medical images: A survey. In *Proceedings of ICRIC 2019*, pages 543–559. Springer, 2020.

[16] I. D. Apostolopoulos and T. A. Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, 43(2):635–640, 2020.

[17] Briseis Aschebrook-Kilfoy, Mary H Ward, Curt T Della Valle, and Melissa C Friesen. Occupation and thyroid cancer. *Occupational and environmental medicine*, 71(5), 2014.

[18] Faiza Asif, Muhammad Riaz Ahmad, and Arshia Majid. Risk factors for thyroid cancer in females using a logit model in lahore, pakistan. *Asian Pacific Journal of Cancer Prevention*, 16(15):6243–6247, 2015.

[19] American Thyroid Association. Thyroid imaging reporting and data systems (tirads) accurately determine the risk of cancer in small thyroid nodules. *Clinical Thyroidology for the Public*, 11:7–8, 2018.

[20] Indriani P Astono, James S Welsh, Christopher W Rowe, and Phillip Jobling. Objective quantification of nerves in immunohistochemistry specimens of thyroid cancer utilising deep learning. *PLOS Computational Biology*, 18(2):e1009912, 2022.

[21] Ghobad Azizi and Carl D. Malchoff. Autoimmune thyroid disease: A risk factor for thyroid cancer. *Endocrine Practice*, 17(2):201–209, 2011. ISSN 1530-891X. doi: https://doi.org/10.4158/EP10123.OR.

[22] Zahra Bahadoran, Parvin Mirmiran, Asghar Ghasemi, Ali Kabir, Fereidoun Azizi, and Farzad Hadaegh. Is dietary nitrate/nitrite exposure a risk factor for development of thyroid abnormality? a systematic review and meta-analysis. *Nitric Oxide*, 47:65–76, 2015.

[23] Ahsan Baidar Bakht, Sajid Javed, Roberto Dina, Hasan Almarzouqi, Ahsan Khandoker, and Naoufel Werghi. Thyroid nodule cell classification in cytology images using transfer learning approach. In *International Conference on Soft Computing and Pattern Recognition*, pages 539–549. Springer, 2020.

[24] Zubair W. Baloch, Seth Fleisher, Virginia A. LiVolsi, and Prabodh K. Gupta. Diagnosis of "follicular neoplasm": a gray zone in thyroid fine-needle aspiration cytology. *Diagnostic cytopathology*, 26(1):41–44, 2002.

[25] Zubair W. Baloch, Sylvia L. Asa, Justine A. Barletta, Ronald A. Ghossein, C. Christofer Juhlin, Chan Kwon Jung, Virginia A. LiVolsi, Mauro G. Papotti, Manuel Sobrinho-Simões, Giovanni Tallini, and Ozgur Mete. Overview of the 2022 who classification of thyroid neoplasms. *Endocrine Pathology*, 33:27–63, 2022. doi: 10.1007/s12022-022-09707-3.

[26] Luigi Barrea, Marco Gallo, Rosaria Maddalena Ruggeri, Paola Di Giacinto, Franz Sesti, Natalie Prinzi, Valerio Adinolfi, Viola Barucca, Valerio Renzelli, Giovanna Muscogiuri, Annamaria Colao, Roberto Baldelli, and on behalf of E.O.L.O. Group. Nutritional status and follicular-derived thyroid cancer: An update. *Critical Reviews in Food Science and Nutrition*, 61(1):25–59, 2021. doi: 10.1080/10408398.2020.1714542.

[27] Bernard Berelson. *Content analysis in communication research*. Free press, 1952.

[28] Vitoantonio Bevilacqua, Antonio Brunetti, Giacomo Donato Cascarano, Flavio Palmieri, Andrea Guerriero, and Marco Moschetta. A deep learning approach for the automatic detection and segmentation in autosomal dominant polycystic kidney disease based on magnetic resonance images. In De-Shuang Huang, Kang-Hyun Jo, and Xiao-Long Zhang, editors, *Intelligent Computing Theories and Application*, pages 643–649, Cham, 2018. Springer International Publishing. ISBN 978-3-319-95933-7.

[29] Dulari Bhatt, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20), 2021. ISSN 2079-9292. doi: 10.3390/electronics10202470.

[30] Pratik Bhowal, Subhankar Sen, Jin Hee Yoon, Zong Woo Geem, and Ram Sarkar. Choquet integral and coalition game-based ensemble of deep learning models for covid-19 screening from chest x-ray images. *IEEE Journal of Biomedical and Health Informatics*, 25(12): 4328–4339, 2021. doi: 10.1109/JBHI.2021.3111415.

[31] Athanasios Bikas and Kenneth D. Burman. Epidemiology of thyroid cancer. *The thyroid and its diseases*, pages 541–547, 2019.

[32] Richard A Bloomfield, Clifford W Welsch, George B Garner, and Merle E Muhrer. Effect of dietary nitrate on thyroid function. *Science*, 134(3491):1690–1690, 1961.

[33] Francoise Borson-Chazot, Sylvain Causeret, Jean-Christophe Lifante, Marylin Augros, Nicole Berger, and Jean-Louis Peix. Predictive factors for recurrence from a series of 74 children and adolescents with differentiated thyroid cancer. *World journal of surgery*, 28(11):1088–1092, 2004.

[34] Keno K. Bressem, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Nature Research Scientific Reports*, 10(1):1–16, 2020.

[35] Anna L. Buczak and Christopher M. Gifford. Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ISI-KDD '10. Association for Computing Machinery, 2010. ISBN 9781450302234. doi: 10.1145/1938606.1938608.

[36] Mateusz Buda, Benjamin Wildman-Tobriner, Jenny K. Hoang, David Thayer, Franklin N. Tessler, William D. Middleton, and Maciej A. Mazurowski. Management of thyroid nodules seen on us images: deep learning may match performance of radiologists. *Radiology*, 292 (3):695 – 701, 2019.

[37] Mateusz Buda, Benjamin Wildman-Tobriner, Kerry Castor, Jenny K. Hoang, and Maciej A. Mazurowski. Deep learning-based segmentation of nodules in thyroid ultrasound: Improving performance by utilizing markers present in the images. *Ultrasound in Medicine & Biology*, 46(2):415–421, 2020. ISSN 0301-5629. doi: https://doi.org/10.1016/j.ultrasmedbio.2019.10.003.

[38] Vijaya Gajanan Buddhavarapu and Angel Arul Jothi J. An experimental study on classification of thyroid histopathology images using transfer learning. *Pattern Recognition Letters*, 140:1–9, 2020. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2020.09.020.

[39] William R. Burns and Martha A. Zeiger. Differentiated thyroid cancer. *Seminars in Oncology*, 37(6):557–566, 2010. ISSN 0093-7754. doi: https://doi.org/10.1053/j.seminoncol.2010.10.008.

[40] Silvio Buscemi, Fatima Maria Massenti, Sonya Vasto, Fabio Galvano, Carola Buscemi, Davide Corleo, Anna Maria Barile, Giuseppe Rosafio, Nadia Rini, and Carla Giordano. Association of obesity and diabetes with thyroid nodules. *Endocrine*, 60(2):339–347, 2018.

[41] Maria E Cabanillas, David G McFadden, and Cosimo Durante. Thyroid cancer. *The Lancet*, 388(10061):2783–2795, 2016. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(16)30172-6.

[42] Kaili Cao and Xiaoli Zhang. An improved res-unet model for tree species classification using airborne high-resolution images. *Remote Sensing*, 12(7), 2020. ISSN 2072-4292. doi: 10.3390/rs12071128.

[43] Ling-Zhi Cao, Xiao-Dong Peng, Jian-Ping Xie, Fan-Hui Yang, Hu-Ling Wen, and Suping Li. The relationship between iodine intake and the risk of thyroid cancer: a meta-analysis. *Medicine*, 96(20), 2017.

[44] Yijuan Cao, Zengyan Wang, Juan Gu, Fangfang Hu, Yujuan Qi, Qianqian Yin, Qingqing Sun, Guotao Li, and Bin Quan. Reproductive factors but not hormonal factors associated with thyroid cancer risk: a systematic review and meta-analysis. *BioMed Research International*, 2015, 2015.

[45] N. Carnes. *Predicting risk of malignancy in patients with indeterminate thyroid nodules.* PhD thesis, Boston University, 2018.

[46] Sunkyung Cha and Sung-Soo Kim. Comorbidity patterns of mood disorders in adult inpatients: Applying association rule mining. *Healthcare*, 9(9), 2021. ISSN 2227-9032. doi: 10.3390/healthcare9091155.

[47] Wai-Kin Chan, Jui-Hung Sun, Miaw-Jene Liou, Yan-Rong Li, Wei-Yu Chou, Feng-Hsuan Liu, Szu-Tah Chen, and Syu-Jyun Peng. Using deep convolutional neural networks for enhanced ultrasonographic image diagnosis of differentiated thyroid cancer. *Biomedicines*, 9(12), 2021. ISSN 2227-9059. doi: 10.3390/biomedicines9121771.

[48] Jamil Ahmed Chandio, Ghulam Ali Mallah, and Noor Ahmed Shaikh. Decision support system for classification medullary thyroid cancer. *IEEE Access*, 8:145216–145226, 2020. doi: 10.1109/ACCESS.2020.3014863.

[49] Chuan-Yu Chang, Shao-Jer Chen, and Ming-Fong Tsai. Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. *Pattern Recognition*, 43(10):3494–3506, 2010. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.04.023.

[50] Yongjun Chang, Anjan Kumar Paul, Namkug Kim, Jung Hwan Baek, Young Jun Choi, Eun Ju Ha, Kang Dae Lee, Hyoung Shin Lee, DaeSeock Shin, and Nakyoung Kim. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. *Medical physics*, 43(1):554–567, 2016.

[51] Phasit Charoenkwan, Wararat Chiangjong, Chanin Nantasenamat, Md Mehedi Hasan, Balachandran Manavalan, and Watshara Shoombuatong. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings in Bioinformatics*, 22(6), 05 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab172.

[52] Jianguo Chen, Kenli Li, Huigui Rong, Kashif Bilal, Nan Yang, and Keqin Li. A disease diagnosis and treatment recommendation system based on big data mining and cloud

computing. *Information Sciences*, 435:124–149, 2018. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2018.01.001.

[53] Yifei Chen, Dandan Li, Xin Zhang, Jing Jin, and Yi Shen. Computer aided diagnosis of thyroid nodules based on the devised small-datasets multi-view ensemble learning. *Medical Image Analysis*, 67:101819, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101819.

[54] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, 6:1–13, 2016.

[55] Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*, 30(4):477–486, 2017.

[56] Ara Cho, Yoosoo Chang, Jiin Ahn, Hocheol Shin, and Seungho Ryu. Cigarette smoking and thyroid cancer risk: a cohort study. *British journal of cancer*, 119(5), 2018.

[57] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE conference on computer vision and pattern recognition*, pages 1251–1258. IEEE, 2017.

[58] Bilal Chouiha and Abdenour Amamra. Thyroid nodules recognition in ultrasound images based on imagenet top-performing deep convolutional neural networks. In *International Conference on Computing Systems and Applications*, pages 313–322. Springer, 2020.

[59] Judy S Crabtree. Fundamentals of heredity. In *Clinical Precision Medicine*, pages 39–47. Elsevier, 2020.

[60] Michael J Crotty. The foundations of social research: Meaning and perspective in the research process. *The foundations of social research*, pages 1–256, 1998.

[61] J. A. Cruz and D. S. Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 2006. ISSN 117693510600200030.

[62] J. Staničić D. Kust and N. Mateša. Bethesda thyroid categories and family history of thyroid disease. *Clinical endocrinology*, 88(3):468–472, 2018. doi: https://doi.org/10.1111/cen.13538.

[63] Rajshree Daulatabad, Roberto Vega, Jacob L. Jaremko, Jeevesh Kapur, Abhilash R. Hareendranathan, and Kumaradeven Punithakumar. Integrating user-input into deep convolutional neural networks for thyroid nodule segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2637–2640, 2021. doi: 10.1109/EMBC46164.2021.9629959.

[64] B D'Avanzo, C La Vecchia, S Franceschi, E Negri, and R Talamini. History of thyroid diseases and subsequent thyroid cancer risk. *Cancer Epidemiology and Prevention Biomarkers*, 4(3):193–199, 1995. ISSN 1055-9965.

[65] Colin M Dayan. Interpretation of thyroid function tests. *The Lancet*, 357(9256):619–624, 2001. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(00)04060-5.

[66] Diana S Dean and Hossein Gharib. Fine-needle aspiration biopsy of the thyroid gland. *Endotext [Internet]*, 2015. URL `https://www.ncbi.nlm.nih.gov/books/NBK285544/`.

[67] T DeCarlo. Paradigms, theories and how they shape a researcher's approach. *Open Social Work Education*, 2018.

[68] S. Deepak and P.M. Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in Biology and Medicine*, 111:103345, 2019. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.103345.

[69] ME Dellavalle and DM Barbano. Iodine content of milk and other foods. *Journal of food protection*, 47(9):678–684, 1984.

[70] Chengwen Deng, Dongyan Han, Ming Feng, Zhongwei Lv, and Dan Li. Differential diagnostic value of the resnet50, random forest, and ds ensemble models for papillary thyroid carcinoma and other thyroid nodules. *Journal of International Medical Research*, 50(4), 2022. doi: 10.1177/03000605221094276.

[71] Jianrui Ding, Zichen Huang, Mengdie Shi, and Chunping Ning. Automatic thyroid ultrasound image segmentation based on u-shaped network. In *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5, 2019. doi: 10.1109/CISP-BMEI48845.2019.8966062.

[72] D. Dov, S. Z. Kovalsky, J. Cohen, D. E. Range, R. Henao, and L. Carin. Thyroid cancer malignancy prediction from whole slide cytopathology images. In *Machine Learning for Healthcare Conference*, pages 553–570. PMLR, 2019.

[73] Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[74] Nguyen Thanh Duc, Yong-Moon Lee, Jae Hyun Park, and Boreom Lee. An ensemble deep learning for automatic prediction of papillary thyroid carcinoma using fine needle aspiration cytology. *Expert Systems with Applications*, 188:115927, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115927.

[75] Cosimo Durante, Teresa Montesano, Massimo Torlontano, Marco Attard, Fabio Monzani, Salvatore Tumino, Giuseppe Costante, Domenico Meringolo, Rocco Bruno, Fabiana Trulli, Michela Massa, Adele Maniglia, Rosaria D'Apollo, Laura Giacomelli, Giuseppe Ronga, Sebastiano Filetti, and on behalf of the PTC Study Group. Papillary thyroid cancer: Time course of recurrences during postsurgery surveillance. *The Journal of Clinical Endocrinology & Metabolism*, 98(2):636–642, 02 2013. ISSN 0021-972X. doi: 10.1210/jc.2012-3401.

[76] Alexey A Efanov, Alina V Brenner, Tetiana I Bogdanova, Lindsey M Kelly, Pengyuan Liu, Mark P Little, Abigail I Wald, Maureen Hatch, Liudmyla Y Zurnadzy, Marina N Nikiforova, Vladimir Drozdovitch, Rebecca Leeman-Neill, Kiyohiko Mabuchi, Mykola D Tronko, Stephen J Chanock, and Yuri E Nikiforov. Investigation of the relationship between radiation dose and gene mutations and fusions in post-chernobyl thyroid cancer. *JNCI: Journal of the National Cancer Institute*, 110(4):371–378, 11 2017. ISSN 0027-8874. doi: 10.1093/jnci/djx209.

[77] Ahmed S. El-Hossiny, Walid Al-Atabany, Osama Hassan, Ahmed Mostafa, and Sherif A. Sami. A robust cnn classification of whole slide thyroid carcinoma images. In *2021 9th In-*

ternational Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC), pages 198–202, 2021. doi: 10.1109/JAC-ECC54461.2021.9691433.

[78] D. D. Elliott Range, D. Dov, S. Z. Kovalsky, R. Henao, L. Carin, and J. Cohen. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. *Cancer Cytopathology*, 128(4):287–295, 2020.

[79] Arthur S Elstein and Alan Schwarz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, 324(7339):729–732, 2002. ISSN 0959-8138. doi: 10.1136/bmj.324.7339.729.

[80] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[81] Rui-Mei Feng, Yi-Nan Zong, Su-Mei Cao, and Rui-Hua Xu. Current cancer situation in china: good or bad news from the 2018 global cancer statistics? *Cancer communications*, 39(1):1–12, 2019.

[82] James J. Figge. Epidemiology of thyroid cancer. *Thyroid Cancer*, 2016.

[83] Maria Fiore, Gea Oliveri Conti, Rosario Caltabiano, Antonino Buffone, Pietro Zuccarello, Livia Cormaci, Matteo Angelo Cannizzaro, and Margherita Ferrante. Role of emerging environmental risk factors in thyroid cancer: A brief review. *International Journal of Environmental Research and Public Health*, 16(7), 2019. ISSN 1660-4601. doi: 10.3390/ijerph16071185.

[84] North American Association for the Study of Obesity, National Heart, Lung, Blood Institute, and NHLBI Obesity Education Initiative. *The practical guide: identification, evaluation, and treatment of overweight and obesity in adults*. National Institutes of Health, National Heart, Lung, and Blood Institute . . . , 2000.

[85] Russ G, Bigorgne C, Royer B, Rouxel A, and M Bienvenu-Perrard. The thyroid imaging reporting and data system (tirads) for ultrasound of the thyroid. *Journal de radiologie*, 92: 701–713, 2011.

[86] Nathaniel L Gage. The paradigm wars and their aftermath a "historical" sketch of research on teaching since 1989. *Educational researcher*, 18(7):4–10, 1989.

[87] Swarnendu Ghosh, Nibaran Das, Ishita Das, and Ujjwal Maulik. Understanding deep learning techniques for image segmentation. *ACM Comput. Surv.*, 52(4), 2019. ISSN 0360-0300. doi: 10.1145/3329784.

[88] Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel. Mutual information analysis. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 426–442. Springer, 2008.

[89] Oliver Gimm. Thyroid cancer. *Cancer Letters*, 163(2):143–156, 2001. ISSN 0304-3835. doi: https://doi.org/10.1016/S0304-3835(00)00697-2.

[90] Salvatore Gitto, Giorgia Grassi, Chiara De Angelis, Cristian Giuseppe Monaco, Silvana Sdao, Francesco Sardanelli, Luca Maria Sconfienza, and Giovanni Mauri. A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound. *La radiologia medica*, 124(2):118–125, 2019.

[91] Elmer Jeto Gomes Ataide, Nikhila Ponugoti, Alfredo Illanes, Simone Schenke, Michael Kreissl, and Michael Friebe. Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. *Sensors*, 20(21), 2020. ISSN 1424-8220. doi: 10.3390/s20216110.

[92] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao abd Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 257–261. IEEE, 2021.

[93] Raymon H. Grogan, Sharone P. Kaplan, Hongyuan Cao, Roy E. Weiss, Leslie J. DeGroot, Cassie A. Simon, Omran M.A. Embia, Peter Angelos, Edwin L. Kaplan, and Rebecca B. Schechter. A study of recurrence and death from papillary thyroid cancer with 27 years of median follow-up. *Surgery*, 154(6):1436–1447, 2013. ISSN 0039-6060. doi: https://doi.org/10.1016/j.surg.2013.07.008.

[94] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725*, 2012.

[95] Qing Guan, Yunjun Wang, Jiajun Du, Yu Qin, Hongtao Lu, Jun Xiang, and Fen Wang. Deep learning based classification of ultrasound images for thyroid nodules: a large scale of pilot study. *Annals of translational medicine*, 7(7), 2019.

[96] Egon G Guba, Yvonna S Lincoln, et al. Competing paradigms in qualitative research. *Handbook of qualitative research*, 2(163-194):105, 1994.

[97] Minghui Guo and Yongzhao Du. Classification of thyroid ultrasound standard plane images using resnet-18 networks. In *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, pages 324–328, 2019. doi: 10.1109/ICASID.2019.8925267.

[98] Aurélien Géron. *Hands on machine learning with Scikit Learn, Keras, and TensorFlow Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[99] Danuta Gąsior-Perczak, Iwona Pałyga, Monika Szymonek, Artur Kowalik, Agnieszka Walczyk, Janusz Kopczyński, Katarzyna Lizis-Kolus, Tomasz Trybek, Estera Mikina, Dorota Szyska-Skrobot, Klaudia Gadawska-Juszczyk, Stefan Hurej, Artur Szczodry, Anna Słuszniak, Janusz Słuszniak, Ryszard Mężyk, Stanisław Góźdź, and Aldona Kowalska. The impact of bmi on clinical progress, response to treatment, and disease course in patients with differentiated thyroid cancer. *PLOS ONE*, 13(10):1–18, 10 2018. doi: 10.1371/journal.pone.0204668.

[100] Martin Halicek, Guolan Lu, James V. Little, Xu Wang, Mihir Patel, Christopher C. Griffith, Mark W. El-Deiry, Amy Y. Chen, and Baowei Fei. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *Journal of biomedical optics*, 22(6):060503, 2017.

[101] Ji Min Han, Tae Yong Kim, Min Ji Jeon, Ji Hye Yim, Won Gu Kim, Dong Eun Song, Suck Joon Hong, Sung Jin Bae, Hong-Kyu Kim, Myung-Hee Shin, Young Kee Shong, and

Won Bae Kim. Obesity is a risk factor for thyroid cancer in a large, ultrasonographically screened population. *Eur J Endocrinol*, 168(6):879–886, 2013.

[102] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12, may 2000. ISSN 0163-5808. doi: 10.1145/335191.335372.

[103] Mi Ah Han and Jin Hwa Kim. Diagnostic x-ray exposure and thyroid cancer risk: Systematic review and meta-analysis. *Thyroid*, 28(2):220–228, 2018. doi: 10.1089/thy.2017.0159.

[104] Steven L. Hancock, I.Ross McDougall, and Louis S. Constine. Thyroid abnormalities after therapeutic external radiation. *International Journal of Radiation OncologyBiologyPhysics*, 31(5):1165–1170, 1995. ISSN 0360-3016. doi: https://doi.org/10.1016/0360-3016(95)00019-U.

[105] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[106] Xiuxiu He, Bang Jun Guo, Yang Lei, Yingzi Liu, Tonghe Wang, Walter J. Curran, Long Jiang Zhang, Tian Liu, and Xiaofeng Yang. 3d thyroid segmentation in ct using self-attention convolutional neural network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 1131445. International Society for Optics and Photonics, 2020. doi: https://doi.org/10.1117/12.2549786.

[107] Kinga Hińcza, Artur Kowalik, and Aldona Kowalska. Current knowledge of germline genetic risk factors for the development of non-medullary thyroid cancer. *Genes*, 10(7), 2019. ISSN 2073-4425. doi: 10.3390/genes10070482.

[108] Yaoshiang Ho and Samuel Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 8:4806–4813, 2020. doi: 10.1109/ACCESS.2019.2962617.

[109] Richard Hodson. *Precision medicine*. Nature, 2016.

[110] Pamela L. Horn-Ross, Alison J. Canchola, Huiyan Ma, Peggy Reynolds, and Leslie Bernstein. Hormonal factors and the risk of papillary thyroid cancer in the california teachers study cohort. *Cancer Epidemiology and Prevention Biomarkers*, 20(8):1751–1759, 2011.

[111] Wenjuan Hu, Hao Wang, Ran Wei, Lanyun Wang, Zedong Dai, Shaofeng Duan, Yaqiong Ge, Pu-Yeh Wu, and Bin Song. Mri-based radiomics analysis to predict preoperative lymph node metastasis in papillary thyroid carcinoma. *Gland Surgery*, 9(5):1214, 2020.

[112] W. Hua, S. Wang, W. Xie, Y. Guo, and X. Jin. Dual-channel convolutional neural network for polarimetric sar images classification. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3201–3204. IEEE, 2019.

[113] Zhu Hua. Identifying research paradigms. *Research methods in intercultural communication: A practical guide*, pages 1–22, 2015.

[114] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[115] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[116] Lulu Huang, Xiuming Feng, Wenjun Yang, Xiangzhi Li, Kang Zhang, Shuzhen Feng, Fei Wang, and Xiaobo Yang. Appraising the effect of potential risk factors on thyroid cancer: A mendelian randomization study. *The Journal of Clinical Endocrinology & Metabolism*, 04 2022. ISSN 0021-972X. doi: 10.1210/clinem/dgac196.

[117] B. Igelnik, Yoh-Han Pao, S.R. LeClair, and Chang Yun Shen. The ensemble approach to neural-network learning and generalization. *IEEE Transactions on Neural Networks*, 10 (1):19–30, 1999. doi: 10.1109/72.737490.

[118] Yasuhiro Ito, Yuri E Nikiforov, Martin Schlumberger, and Riccardo Vigneri. Increasing incidence of thyroid cancer: controversies explored. *Nature Reviews Endocrinology*, 9(3): 178–184, 2013.

[119] M. Jajroudi, T. Baniasadi, L. Kamkar, F. Arbabi, M. Sanei, and M. Ahmadzade. Prediction of survival in thyroid cancer using data mining technique. *Technology in Cancer Research & Treatment*, 13(4):353–359, 2014. doi: 10.7785/tcrt.2012.500384.

[120] O. Jamsheela. Analysis of association among various attributes in medical data of heart patients by using data mining methods. *International Journal of Applied Science and Engineering*, 18:1–6, June 2021. ISSN 1727-7841. doi: 10.6703/IJASE.202106_18(2).009.

[121] Sabrina Jegerlehner, Jean-Luc Bulliard, Drahomir Aujesky, Nicolas Rodondi, Simon Germann, Isabelle Konzelmann, Arnaud Chiolero, and NICER Working Group. Overdiagnosis and overtreatment of thyroid cancer: A population-based temporal trend study. *PloS one*, 12(6), 2017. ISSN e0179387.

[122] E. Y. Jeong, H. L. Kim, E. J. Ha, S. Y. Park, Y. J. Cho, and M. Han. Computer-aided diagnosis system for thyroid nodules on ultrasonography: diagnostic performance and reproducibility based on the experience level of operators. *European radiology*, 29(4): 1978–1985, 2019.

[123] D. Jimenez. Dynamically weighted ensemble neural networks for classification. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, volume 1, pages 753–756 vol.1, 1998. doi: 10.1109/IJCNN.1998.682375.

[124] Choi Wook Jin and Kim Jeongseon. Dietary factors and the risk of thyroid cancer: A review. *Clinical Nutrition Research*, 3(2):75–88, 2014. doi: 10.7762/cnr.2014.3.2.75.

[125] Weiqiu Jin, Shuqin Dong, Changzi Dong, and Xiaodan Ye. Hybrid ensemble model for differential diagnosis between covid-19 and common viral pneumonia by chest x-ray radiograph. *Computers in Biology and Medicine*, 131:104252, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104252.

[126] Su jin Kim, Seog Yun Park, You Jin Lee, Eun Kyung Lee, Seok ki Kim, Tae Hyun Kim, Yu-Seog Jung, Junsun Ryu, Jun Pyo Myong, and Ki-Wook Chung. Risk factors for recurrence

after therapeutic lateral neck dissection for primary papillary thyroid cancer. *Annals of surgical oncology*, 21(6):1884–1890, 2014.

[127] Paul Johannesson and Erik Perjons. Research paradigms. In *An Introduction to Design Science*, pages 167–179. Springer, 2014.

[128] JR. John D. Boice. Ionizing radiation. In David Schottenfeld and Joseph F. Fraumeni, editors, *Cancer epidemiology and prevention*, chapter 15, page 259. Oxford University Press, Oxford, 2006.

[129] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[130] R. Burke Johnson and Anthony J. Onwuegbuzie. Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7):14–26, 2004. doi: 10.3102/0013189X033007014.

[131] Jacqueline Jonklaas, Mark Danielsen, and Hong Wang. A pilot study of serum selenium, vitamin d, and thyrotropin concentrations in patients with thyroid cancer. *Thyroid*, 23(9): 1079–1086, 2013.

[132] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

[133] Minakshi Kaushik, Rahul Sharma, Sijo Arakkal Peious, Mahtab Shahin, Sadok Ben Yahia, and Dirk Draheim. A systematic assessment of numerical association rule mining methods. *SN Computer Science*, 2(5):1–13, 2021.

[134] F. Khodamoradi, M. Ghoncheh, A. Mehri, S. Hassanipour, and H. Salehiniya. Incidence, mortality, and risk factors of thyroid cancer in the world: a review. *World Cancer Research Journal*, 5(2), 2018.

[135] E. Kim, M. Corte-Real, and Z. Baloch. A deep semantic mobile application for thyroid cytopathology. *Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations*, 9789:97890A, 2016.

[136] Jina Kim, Jessica E Gosnell, and Sanziana A Roman. Geographic influences in the global rise of thyroid cancer. *Nature Reviews Endocrinology*, 16(1):17–29, 2020.

[137] Kyungsik Kim, Sun Wook Cho, Young Joo Park, Kyu Eun Lee, Dong-Wook Lee, and Sue K. Park. Association between iodine intake, thyroid function, and papillary thyroid cancer: A case-control study. *Endocrinology and metabolism*, 36(4), 2021. doi: 10.3803/EnM.2021.1034.

[138] Min Jhi Kim, Daham Kim, Ja Seung Koo, Ju Hee Lee, and Kee-Hyun Nam. Vitamin d receptor expression and its clinical significance in papillary thyroid cancer. *Technology in Cancer Research & Treatment*, 21:15330338221089933, 2022. doi: 10.1177/15330338221089933.

[139] S. K. Kim, J. W. Woo, I. Park, J. H. Lee, J. H. Choe, J. H. Kim, and J. S Kim. Computed tomography-detected central lymph node metastasis in ultrasonography node-negative papillary thyroid carcinoma: Is it really significant? *Annals of surgical oncology*, 24(2):442–449, 2017.

[140] Yeon-Jae Kim, Yangsean Choi, Su-Jin Hur, Ki-Sun Park, Hyun-Jin Kim, Minkook Seo, Min Kyoung Lee, So-Lyung Jung, and Chan Kwon Jung. Deep convolutional neural network for classification of thyroid nodules on ultrasound: Comparison of the diagnostic performance with that of radiologists. *European Journal of Radiology*, 152:110335, 2022.

[141] Cari M. Kitahara and Julie A. Sosa. Understanding the ever-changing incidence of thyroid cancer. *Nature Reviews Endocrinology*, 16(11):617–618, 2020. doi: https://doi.org/10.1038/s41574-020-00414-9.

[142] Cari M Kitahara, Elizabeth A Platz, Laura E Beane Freeman, Ann W Hsing, Martha S Linet, Yikyung Park, Catherine Schairer, Arthur Schatzkin, James M Shikany, and Amy Berrington de González. Obesity and thyroid cancer risk among us men and women: a

pooled analysis of five prospective studies. *Cancer epidemiology, biomarkers & prevention*, 20(3):464–472, 2011.

[143] Cari M Kitahara, Dóra Kormendiné Farkas, Jens Otto L Jørgensen, Deirdre Cronin-Fenton, and Henrik Toft Sørensen. Benign thyroid diseases and risk of thyroid cancer: A nationwide cohort study. *The Journal of Clinical Endocrinology & Metabolism*, 103(6):2216–2224, 03 2018. ISSN 0021-972X. doi: 10.1210/jc.2017-02599.

[144] Gregory Kline and Hossein Sadrzadeh. Chapter 2 - thyroid disorders. In Hossein Sadrzadeh and Gregory Kline, editors, *Endocrine Biomarkers*, pages 41–93. Elsevier, 2017. ISBN 978-0-12-803412-5. doi: https://doi.org/10.1016/B978-0-12-803412-5.00002-1.

[145] P Kmieć and K Sworczak. Vitamin d in thyroid disorders. *Experimental and Clinical Endocrinology & Diabetes*, 123(07):386–393, 2015.

[146] Su Yeon Ko, Ji Hye Lee, Jung Hyun Yoon, Hyesun Na, Eunhye Hong, Kyunghwa Han, Inkyung Jung, Eun-Kyung Kim, Hee Jung Moon, Vivian Y. Park, Eunjung Lee, and Jin Young Kwak. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head & neck*, 41(4):885–891, 2019.

[147] Matthieu Komorowski, Dominic C Marshall, Justin D Salciccioli, and Yves Crutain. Exploratory data analysis. *Secondary analysis of electronic health records*, pages 185–203, 2016.

[148] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001. ISSN 0933-3657. doi: https://doi.org/10.1016/S0933-3657(01)00077-X.

[149] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *In: Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[150] Viksit Kumar, Jeremy Webb, Adriana Gregory, Duane D. Meixner, John M. Knudsen, Matthew Callstrom, Mostafa Fatemi, and Azra Alizad. Automated segmentation of thyroid

nodule, gland, and cystic components from ultrasound images using deep learning. *IEEE Access*, 8:63482–63496, 2020. doi: 10.1109/ACCESS.2020.2982390.

[151] Hyemi Kwon, Yoosoo Chang, Ara Cho, Jiin Ahn, Se Eun Park, Cheol-Young Park, Won-Young Lee, Ki-Won Oh, Sung-Woo Park, Hocheol Shin, Seungho Ryu, and Eun-Jung Rhee. Metabolic obesity phenotypes and thyroid cancer risk: A cohort study. *Thyroid*, 29(3): 349–358, 2019. doi: 10.1089/thy.2018.0327.

[152] A. Kyriacou, V. Tziaferi, and M. Toumba. Stress, thyroid dysregulation and thyroid cancer in children and adolescents: Proposed impending mechanisms. *Hormone Research in Paediatrics*, 2022. doi: https://doi.org/10.1159/000524477.

[153] Nathan Laney, Jane Meza, Elizabeth Lyden, Judi Erickson, Kelly Treude, and Whitney Goldner. The prevalence of vitamin d deficiency is similar between thyroid nodule and thyroid cancer patients. *International journal of endocrinology*, 2010, 2010.

[154] R Larisch, K Kley, S Nikolaus, W Sitte, M Franz, H Hautzel, W Tress, and H-W Müller. Depression and anxiety in different thyroid function states. *Hormone and metabolic research*, 36(09):650–653, 2004.

[155] Sophie Leboulleux, Isabelle Borget, and Martin Schlumberger. Post-operative radioactive iodine administration in patients with low-risk thyroid cancer. *Nature Reviews Endocrinology*, 2022. doi: 10.1038/s41574-022-00709-z.

[156] Yann LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[157] Han Na Lee, Ji Young An, Kyung Mi Lee, Eui Jong Kim, Woo Suk Choi, and Deog Yoon Kim. Salivary gland dysfunction after radioactive iodine (i-131) therapy in patients following total thyroidectomy: emphasis on radioactive iodine therapy dose. *Clinical Imaging*, 39 (3):396–400, 2015. ISSN 0899-7071. doi: https://doi.org/10.1016/j.clinimag.2014.12.018.

[158] Jeong Hoon Lee, Eun Ju Ha, and Ju Han Kim. Application of deep learning to the diagnosis

of cervical lymph node metastasis from thyroid cancer with ct. *European radiology*, 29(10): 5452–5457, 2019.

[159] Su Jung Lee and Kathleen B. Cartmell. An association rule mining analysis of lifestyle behavioral risk factors in cancer survivors with high cardiovascular disease risk. *Journal of Personalized Medicine*, 11(5), 2021. ISSN 2075-4426. doi: 10.3390/jpm11050366.

[160] C Leux and P Guenel. Risk factors of thyroid tumors: role of environmental and occupational exposures to chemical pollutants. *Revue d'epidemiologie et de sante publique*, 58(5): 359–367, 2010.

[161] Dongyang Li, Dan Yang, and Jing Zhang. Arb: Knowledge discovery and disease diagnosis on thyroid disease diagnosis integrating association rule with bagging algorithm. *Engineering Letters*, 28(2), 2020.

[162] Guannan Li, Mingxia Liu, Quansen Sun, Dinggang Shen, and Li Wang. Early diagnosis of autism disease by multi-channel cnns. In Yinghuan Shi, Heung-Il Suk, and Mingxia Liu, editors, *Machine Learning in Medical Imaging*, pages 303–309, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00919-9.

[163] Wenjun Li, Siyi Cheng, Kai Qian, Keqiang Yue, and Hao Liu. Automatic recognition and classification system of thyroid nodules in ct images based on cnn. *Computational Intelligence and Neuroscience*, 2021.

[164] Xiangchun Li, Sheng Zhang, Qiang Zhang, Xi Wei, Yi Pan, Jing Zhao, Xiaojie Xin, Chunxin Qin, Xiaoqing Wang, Jianxin Li, Fan Yang, Yanhui Zhao, Meng Yang, Qinghua Wang, Zhiming Zheng, Xiangqian Zheng, Xiangming Yang, Christopher T Whitlow, Metin Nafi Gurcan, Lun Zhang, Xudong Wang, Boris C Pasche, Ming Gao, Wei Zhang, and Kexin Chen. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology*, 20(2):193–201, 2019. ISSN 1470-2045. doi: https://doi.org/10.1016/S1470-2045(18)30762-9.

[165] Xuewei Li, Shuaijie Wang, Xi Wei, Jialin Zhu, Ruiguo Yu, Mankun Zhao, Mei Yu, Zhiqiang Liu, and Shupei Liu. Fully convolutional networks for ultrasound image segmentation of thyroid nodules. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 886–890, 2018. doi: 10.1109/HPCC/SmartCity/DSS.2018.00147.

[166] X. Liang, J. Yu, J. Liao, and Z. Chen. Convolutional neural network for breast and thyroid nodules diagnosis in ultrasound imaging. *BioMed Research International*, 2020.

[167] Zhaohui Liang, Andrew Powell, Ilker Ersoy, Mahdieh Poostchi, Kamolrat Silamut, Kannappan Palaniappan, Peng Guo, Md Amir Hossain, Antani Sameer, Richard James Maude, Jimmy Xiangji Huang, Stefan Jaeger, and George Thoma. Cnn-based image analysis for malaria diagnosis. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 493–496, 2016. doi: 10.1109/BIBM.2016.7822567.

[168] Kyoung Ja Lim, Chul Soon Choi, Dae Young Yoon, Suk Ki Chang, Kwang Ki Kim, Heon Han, Sam Soo Kim, Jiwon Lee, and Yong Hwan Jeon. Computer-aided diagnosis for the differentiation of malignant from benign thyroid nodules on ultrasonography. *Academic Radiology*, 15(7):853–858, 2008. ISSN 1076-6332. doi: https://doi.org/10.1016/j.acra.2007.12.022.

[169] James W. Little. Thyroid disorders. part iii: neoplastic thyroid disease. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 102(3):275–280, 2006. ISSN 1079-2104. doi: https://doi.org/10.1016/j.tripleo.2005.05.071.

[170] Huan Liu, Hongjun Lu, Ling Feng, and Farhad Hussain. Efficient search of reliable exceptions. In Ning Zhong and Lizhu Zhou, editors, *Methodologies for Knowledge Discovery and Data Mining*, pages 194–204, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48912-2.

[171] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Deep multi-task multi-channel learning for joint classification and regression of brain status. In *International conference*

on medical image computing and computer-assisted intervention, pages 3–11, Cham, 2017. Springer.

[172] Tao Liu, Yun Tian, Shifeng Zhao, Xiaoying Huang, and Qingjun Wang. Residual convolutional neural network for cardiac image segmentation and heart disease diagnosis. *IEEE Access*, 8:82153–82161, 2020. doi: 10.1109/ACCESS.2020.2991424.

[173] Tianjiao Liu, Shuaining Xie, Jing Yu, Lijuan Niu, and Weidong Sun. Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 919–923, 2017. doi: 10.1109/ICASSP.2017.7952290.

[174] Tianjiao Liu, Qianqian Guo, Chunfeng Lian, Xuhua Ren, Shujun Liang, Jing Yu, Lijuan Niu, Weidong Sun, and Dinggang Shen. Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. *Medical Image Analysis*, 58:101555, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101555.

[175] Y. I. Liu, A. Kamaya, T. S. Desser, and D. L. Rubin. A bayesian classifier for differentiating benign versus malignant thyroid nodules using sonographic features. In *AMIA Annual Symposium Proceedings*, page 419. American Medical Informatics Association, 2008.

[176] Yihao Liu, Lei Su, and Haipeng Xiao. Review of factors related to the thyroid cancer epidemic. *International journal of endocrinology*, 2017, 2017. doi: https://doi.org/10.1155/2017/5308635.

[177] Yueyi I. Liu, Aya Kamaya, Terry S. Desser, and Daniel L. Rubin. A bayesian classifier for differentiating benign versus malignant thyroid nodules using sonographic features. In *AMIA Annual Symposium Proceedings*, volume 2008, page 419, 2008.

[178] Andrea Loddo, Sara Buttau, and Cecilia Di Ruberto. Deep learning based pipelines for alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method. *Computers in Biology and Medicine*, 141:105032, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.105032.

[179] Jintao Lu, Xi Ouyang, Xueda Shen, Tianjiao Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. Gan-guided deformable attention network for identifying thyroid nodules in ultrasound images. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1582–1590, 2022. doi: 10.1109/JBHI.2022.3153559.

[180] Wei Lu, Lianzhen Zhong, Di Dong, Mengjie Fang, Qi Dai, Shaoyi Leng, Liwen Zhang, Wei Sun, Jie Tian, Jianjun Zheng, and Yinhua Jin. Radiomic analysis for preoperative prediction of cervical lymph node metastasis in patients with papillary thyroid carcinoma. *European Journal of Radiology*, 118:231–238, 2019. ISSN 0720-048X. doi: https://doi.org/10.1016/j.ejrad.2019.07.018.

[181] Yu Lun, Xiaoyu Wu, Qian Xia, Yanshuo Han, Xiaoyu Zhang, Zhimin Liu, Fengyi Wang, Zhiquan Duan, Shijie Xin, and Jian Zhang. Hashimoto's thyroiditis as a risk factor of papillary thyroid cancer may improve cancer prognosis. *Otolaryngology–Head and Neck Surgery*, 148(3):396–402, 2013. doi: 10.1177/0194599812472426.

[182] Juhua Luo, Michael Hendryx, JoAnn E. Manson, XiaoYun Liang, and Karen L. Margolis. Hysterectomy, oophorectomy, and risk of thyroid cancer. *The Journal of Clinical Endocrinology & Metabolism*, 101(10):3812–3819, 07 2016. ISSN 0021-972X. doi: 10.1210/jc.2016-2011.

[183] Fenglong Ma, Muchao Ye, Junyu Luo, Cao Xiao, and Jimeng Sun. Advances in mining heterogeneous healthcare data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 4050–4051. Association for Computing Machinery, 2021. ISBN 9781450383325.

[184] Jie Ma, Min Huang, Li Wang, Wei Ye, Yan Tong, and Hanmin Wang. Obesity and risk of thyroid cancer: evidence from a meta-analysis of 21 observational studies. *Medical science monitor: international medical journal of experimental and clinical research*, 21:283, 2015. doi: 10.12659/MSM.892035.

[185] Jie Ma, Min Huang, Li Wang, Wei Ye, Yan Tong, and Hanmin Wang. Obesity and risk of thyroid cancer: evidence from a meta-analysis of 21 observational studies. *Medical science*

monitor : international medical journal of experimental and clinical research, 21:283–291, 2015. doi: 10.12659/MSM.892035. URL `https://doi.org/10.12659/MSM.892035`.

[186] Jinlian Ma, Fa Wu, Tian'an Jiang, Qiyu Zhao, and Dexing Kong. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *International journal of computer assisted radiology and surgery*, 12(11):1895–1910, 2017.

[187] Jinlian Ma, Fa Wu, Tian'an Jiang, Qiyu Zhao, and Dexing Kong. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *International journal of computer assisted radiology and surgery*, 12(11):1895–1910, 2017.

[188] Jinlian Ma, Fa Wu, Jiang Zhu, Dong Xu, and Dexing Kong. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. In *Ultrasonics*, pages 221–230. Elsevier, 2017.

[189] L. Ma, C. Ma, Y. Liu, and X. Wang. Thyroid diagnosis from spect images using convolutional neural network with optimization. *Computational intelligence and neuroscience*, 2019.

[190] Wendy J Mack, Susan Preston-Martin, Leslie Bernstein, and Dajun Qian. Lifestyle and other risk factors for thyroid cancer in los angeles county females. *Annals of Epidemiology*, 12(6):395–401, 2002. ISSN 1047-2797. doi: https://doi.org/10.1016/S1047-2797(01)00281-2.

[191] A. Maier, C. Syben, T. Lasser, and C. Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29:86–101, 2019.

[192] Arathy S. Mannathazhathu, Preethi S. George, Sreekala Sudhakaran, Durga Vasudevan, Jagathnath Krishna KM, Christopher Booth, and Aleyamma Mathew. Reproductive factors and thyroid cancer risk: Meta-analysis. *Head & neck*, 41(12):4199–4208, 2019. doi: https://doi.org/10.1002/hed.25945.

[193] Marjory Alana Marcello, Aline Castaldi Sampaio, Bruno Geloneze, Ana Carolina Junqueira Vasques, Ligia Vera Montalli Assumpção, and Laura Sterian Ward. Obesity and excess

protein and carbohydrate consumption are risk factors for thyroid cancer. *Nutrition and cancer*, 64(8):1190–1195, 2012.

[194] Marjory Alana Marcello, Lucas Leite Cunha, Fernando Assis Batista, and Laura Sterian Ward. Obesity and thyroid cancer. *Endocrine-related cancer*, 21(5):T255–T271, 2014.

[195] Gonçalo Marques, Deevyankar Agarwal, and Isabel de la Torre Díez. Automated medical diagnosis of covid-19 through efficientnet convolutional neural network. *Applied Soft Computing*, 96:106691, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106691.

[196] T. Martín-Noguerol, F. Paulano-Godino, R. López-Ortega, J.M. Górriz, R.F. Riascos, and A. Luna. Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. *Clinical Radiology*, 76(5): 317–324, 2021. ISSN 0009-9260. doi: https://doi.org/10.1016/j.crad.2020.11.113.

[197] T. Masuda, T. Nakaura, Y. Funama, K. Sugino, T. Sato, T. Yoshiura, Y. Baba, and K. Awai. Machine learning to identify lymph node metastasis from thyroid cancer in patients undergoing contrast-enhanced ct studies. *Radiography*, 27(3):920–926, 2021. ISSN 1078-8174. doi: https://doi.org/10.1016/j.radi.2021.03.001.

[198] Antonio Matrone, Federica Ferrari, Ferruccio Santini, and Rossella Elisei. Obesity as a risk factor for thyroid cancer. *Current Opinion in Endocrinology, Diabetes and Obesity*, 27 (5):358–363, 2020. doi: 10.1097/MED.0000000000000556.

[199] Ernest L. Mazzaferri and Sissy M. Jhiang. Long-term impact of initial surgical and medical therapy on papillary and follicular thyroid cancer. *The American Journal of Medicine*, 97 (5):418–428, 1994. ISSN 0002-9343. doi: https://doi.org/10.1016/0002-9343(94)90321-2.

[200] Patrick E McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1, 2010.

[201] C.A. Meier. Role of imaging in thyroid disease. In *Diseases of the Brain, Head & Neck, Spine*, pages 243–250. Springer Milan, 2008.

[202] Cari L. Meinhold, Elaine Ron, Sara J. Schonfeld, Bruce H. Alexander, D. Michal Freedman, Martha S. Linet, and Amy Berrington de González. Nonradiation risk factors for thyroid cancer in the us radiologic technologists study. *American Journal of Epidemiology*, 171(2): 242–252, 11 2009. ISSN 0002-9262. doi: 10.1093/aje/kwp354.

[203] Cari L Meinhold, Elaine Ron, Sara J Schonfeld, Bruce H Alexander, D Michal Freedman, Martha S Linet, and Amy Berrington de González. Nonradiation risk factors for thyroid cancer in the us radiologic technologists study. *American journal of epidemiology*, 171(2): 242–252, 2010.

[204] MD. Melissa, C.S. Thyroid cancer symptoms, signs, treatment, types & survival rate, 2020.

[205] Takehiro Michikawa, Manami Inoue, Taichi Shimazu, Norie Sawada, Motoki Iwasaki, Shizuka Sasazuki, Taiki Yamaji, and Shoichiro Tsugane. Seaweed consumption and the risk of thyroid cancer in women. *European journal of cancer prevention*, 21(3):254–260, 2012.

[206] Magdalena Mileva1, Bojana Stoilovska, Anamarija Jovanovska, Ana Ugrinska, Gordana Petrushevska, Slavica Kostadinova-Kunovska, Daniela Miladinova, and Venjamin Majstorov. Thyroid cancer detection rate and associated risk factors in patients with thyroid nodules classified as bethesda category iii. *Radiology and oncology*, 52(4):370, 2018.

[207] Yasir Iqbal Mir. *Improved Thyroid Disease Prediction Model Using Data Mining Techniques with Outlier Detection*, pages 129–161. Springer International Publishing, Cham, 2021. ISBN 978-3-030-71975-3. doi: 10.1007/978-3-030-71975-3_5.

[208] Nia S Mitchell, Victoria A Catenacci, Holly R Wyatt, and James O Hill. Obesity: overview of an epidemic. *Psychiatric clinics*, 34(4):717–732, 2011.

[209] Woo Kyung Moon, Yan-Wei Lee, Hao-Hsiang Ke, Su Hyun Lee, Chiun-Sheng Huang, and Ruey-Feng Chang. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 190:105361, 2020. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105361.

[210] Joanna Moriarty. *Qualitative methods overview.* National Institute for Health Research School for Social Care, 2011.

[211] Moustafa Mourad, Sami Moubayed, Aaron Dezube, Youssef Mourad, Kyle Park, Albertina Torreblanca-Zanca, José S. Torrecilla, John C. Cancilla, and Jiwu Wang. Machine learning and feature selection applied to seer data to reliably assess thyroid cancer prognosis. *Scientific Reports*, 10(1):1–11, 2020.

[212] Hatwib Mugasa, Sumeet Dua, Joel E.W. Koh, Yuki Hagiwara, Oh Shu Lih, Chakri Madla, Pailin Kongmebhol, Kwan Hoong Ng, and U. Rajendra Acharya. An adaptive feature extraction model for classification of thyroid lesions in ultrasound images. *Pattern Recognition Letters*, 131:463–473, 2020. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2020.02.009.

[213] Rashmi Mullur, Yan-Yun Liu, and Gregory A. Brent. Thyroid hormone regulation of metabolism. *Physiological Reviews*, 94(2):355–382, 2014. doi: 10.1152/physrev.00030.2013.

[214] G Muscogiuri, G Tirabassi, G Bizzaro, F Orio, SA Paschou, A Vryonidou, G Balercia, Y Shoenfeld, and A Colao. Vitamin d and thyroid disease: to d or not to d? *European journal of clinical nutrition*, 69(3):291–296, 2015.

[215] Samreen Naeem, Aqib Ali, Salman Qadri, Wali Khan Mashwani, Nasser Tairan, Habib Shah, Muhammad Fayaz, Farrukh Jamal, Christophe Chesneau, and Sania Anam. Machine-learning based hybrid-feature analysis for liver cancer classification using fused (mr and ct) images. *Applied Sciences*, 10(9), 2020. ISSN 2076-3417. doi: 10.3390/app10093134.

[216] Yuji Nagayama. Thyroid autoimmunity and thyroid cancer–the pathogenic connection: a 2018 update. *Hormone and Metabolic Research*, 50(12):922–931, 2018. doi: 10.1055/a-0648-4593.

[217] Ahmed Naglah, Fahmi Khalifa, Reem Khaled, Ahmed Abdel khalek Abdel Razek, and Ayman El-Baz. Thyroid cancer computer-aided diagnosis system using mri-based multi-

input cnn model. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1691–1694, 2021. doi: 10.1109/ISBI48211.2021.9433841.

[218] Nikhil S. Narayan, Pina Marziliano, and Christopher G.L. Hobbs. Automatic removal of manually induced artefacts in ultrasound images of thyroid gland. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3399–3402. IEEE, 2013.

[219] Isadore Newman, Carolyn R Benz, and Carolyn S Ridenour. *Qualitative-quantitative research methodology: Exploring the interactive continuum*. SIU Press, 1998.

[220] Dat Tien Nguyen, Tuyen Danh Pham, Ganbayar Batchuluun, Hyo Sik Yoon, and Kang Ryoung Park. Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains. *Journal of Clinical Medicine*, 8(11), 2019. ISSN 2077-0383. doi: 10.3390/jcm8111976.

[221] Dat Tien Nguyen, Jin Kyu Kang, Tuyen Danh Pham, Ganbayar Batchuluun, and Kang Ryoung Park. Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. *MDPI*, 20(7):1822, 2020.

[222] M. Niepert, M. Ahmed, and K. Kutzkov. Learning convolutional neural networks for graphs. In *International Conference on Machine Learning*, pages 2014–2023. PMLR, 2016.

[223] Yuri E Nikiforov and Marina N Nikiforova. Molecular genetics and diagnosis of thyroid cancer. *Nature Reviews Endocrinology*, 7(10):569–580, 2011.

[224] P. Nix, A. Nicolaides, and A. P. Coatesworth. Thyroid cancer review 1: presentation and investigation of thyroid cancer. *International journal of clinical practice*, 59(11):1340–1344, 2005.

[225] Salem I. Noureldine and Ralph P. Tufano. Association of hashimoto's thyroiditis and thyroid cancer. *Current Opinion in Oncology*, 27(1):21–25, 2015. doi: 10.1097/CCO.0000000000000150.

[226] Anan Nugroho, Risanuri Hidayat, and Hanung Adi Nugroho. Artifact removal in radiological ultrasound images using selective and adaptive median filter. In *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, ICCSP '19, page 237–241, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366182. doi: 10.1145/3309074.3309119.

[227] Hanung Adi Nugroho and Eka Legya Frannita. Impact of implementing data balancing method in intelligent thyroid cancer detection. In *2021 International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, pages 102–106, 2021. doi: 10.1109/COSITE52651.2021.9649624.

[228] Hanung Adi Nugroho and Eka Legya Frannita. Thyroid cancer classification using transfer learning. In *2021 International Conference on Computer Science and Engineering (IC2SE)*, volume 1, pages 1–5, 2021. doi: 10.1109/IC2SE52832.2021.9791905.

[229] Hanung Adi Nugroho, Eka Legya Frannita, Anan Nugroho, Zulfanahri, Igi Ardiyanto, and Lina Choridah. Classification of thyroid nodules based on analysis of margin characteristic. In *2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 47–51, 2017. doi: 10.1109/IC3INA.2017.8251738.

[230] Hanung Adi Nugroho, Eka Legya Frannita, and Augustine Herini Tita Hutami. Thyroid nodules categorization based on margin features using deep learning. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 499–504, 2020. doi: 10.1109/ISRITI51436.2020.9315329.

[231] Hanung Adi Nugroho, Eka Legya Frannita, and Rizki Nurfauzi. An automated detection and segmentation of thyroid nodules using res-unet. In *2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 181–185, 2021. doi: 10.23919/EECSI53397.2021.9624248.

[232] Hanung Adi Nugroho, Zulfanahri, Eka Legya Frannita, Igi Ardiyanto, and Lina Choridah. Computer aided diagnosis for thyroid cancer system based on internal and external characteristics. *Journal of King Saud University - Computer and Information Sciences*, 33(3): 329–339, 2021. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2019.01.007.

[233] Hanung Adi Nugroho, Zulfanahri, Eka Legya Frannita, Igi Ardiyanto, and Lina Choridah. Computer aided diagnosis for thyroid cancer system based on internal and external characteristics. *Journal of King Saud University - Computer and Information Sciences*, 33(3): 329–339, 2021. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2019.01.007.

[234] Fusheng Ouyang, Baoliang Guo, Lizhu Ouyang, Ziwei Liu, Shaojia Lin, Wei Meng, Xiyi Huang, Haixiong Chen, Qiugen Hu, and Shaoming Yang. Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. *European Journal of Radiology*, 113:251–257, 2019.

[235] Kyoungjune Pak, Sunghwan Suh, Seong Jang Kim, and In-Joo Kim. Prognostic value of genetic mutations in thyroid cancer: A meta-analysis. *Thyroid*, 25(1):63–70, 2015. doi: 10.1089/thy.2014.0241.

[236] E.I. Papageorgiou, C.D. Stylios, and P.P. Groumpos. An integrated two-level hierarchical system for decision making in radiation therapy based on fuzzy cognitive maps. *IEEE Transactions on Biomedical Engineering*, 50(12):1326–1339, 2003. doi: 10.1109/TBME.2003.819845.

[237] Elpiniki Papageorgiou, Chrysostomos Stylios, and Peter Groumpos. A combined fuzzy cognitive map and decision trees model for medical decision making. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6117–6120, 2006. doi: 10.1109/IEMBS.2006.260354.

[238] Anastasios Papanastasiou, Konstantinos Sapalidis, Dimitrios G. Goulis, Nikolaos Michalopoulos, Evangelia Mareti, Stylianos Mantalovas, and Isaak Kesisoglou. Thyroid nodules as a risk factor for thyroid cancer in patients with graves' disease: A systematic review and meta-analysis of observational studies in surgically treated patients. *Clinical Endocrinology*, 91(4):571–577, 2019.

[239] Theodora Pappa and Maria Alevizaki. Obesity and thyroid cancer: A clinical update. *Thyroid*, 24(2):190–199, 2014. doi: 10.1089/thy.2013.0232.

[240] Young Min Park and Byung-Joo Lee. Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence. *Scientific Reports*, 11 (1):1–7, 2021.

[241] Michael Quinn Patton. *Qualitative research & evaluation methods*. sage, 2002.

[242] S Pavithra, G Yamuna, and R Arunkumar. Deep learning method for classifying thyroid nodules using ultrasound images. In *2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, pages 1–6, 2022. doi: 10.1109/ICSTSN53084.2022.9761364.

[243] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero. An open access thyroid ultrasound image database. In *International Symposium on Medical Information Processing and Analysis*, page 9287. International Society for Optics and Photonics, 2015.

[244] Maria Rosa Pelizzo, Isabella Merante Boschin, Antonio Toniato, Andrea Piotto, Costantino Pagetta, Milton D. Gross, Adil FRCP Al-Nahhas, and Domenico Rubello. Papillary thyroid carcinoma: 35-year outcome and prognostic factors in 1858 patients. *Clinical nuclear medicine*, 32(6):440–444, 2007.

[245] Gabriella Pellegriti, Francesco Frasca, Concetto Regalbuto, Sebastiano Squatrito, and Riccardo Vigneri. Worldwide increasing incidence of thyroid cancer: update on epidemiology and risk factors. *Journal of cancer epidemiology*, 2013.

[246] Emily Peterson, Prithwish De, and Robert Nuttall. Bmi, diet and female reproductive factors as risks for thyroid cancer: A systematic review. *PLOS ONE*, 7(1):1–10, 01 2012. doi: 10.1371/journal.pone.0029177.

[247] Margherita Pizzato, Mengmeng Li, Jerome Vignat, Mathieu Laversanne, Deependra Singh, Carlo La Vecchia, and Salvatore Vaccarella. The epidemiological landscape of thyroid cancer worldwide: Globocan estimates for incidence and mortality rates in 2020. *The Lancet Diabetes & Endocrinology*, 10(4):264–272, 2022. ISSN 2213-8587. doi: https://doi.org/10.1016/S2213-8587(22)00035-3.

[248] Robin L Plackett. Karl pearson and the chi-squared test. *International statistical review/revue internationale de statistique*, pages 59–72, 1983.

[249] Prabal Poudel, Alfredo Illanes, Debdoot Sheet, and Michael Friebe. Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches. *Journal of healthcare engineering*, 2018, 2018.

[250] Prabal Poudel, Alfredo Illanes, Debdoot Sheet, and Michael Friebe. Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches. *Journal of healthcare engineering*, 2018, 2018. doi: https://doi.org/10.1155/2018/8087624.

[251] Swarnalatha Purushotham and B. K. Tripathy. Evaluation of classifier models using stratified tenfold cross validation techniques. In P. Venkata Krishna, M. Rajasekhara Babu, and Ezendu Ariwa, editors, *Global Trends in Information Systems and Software Applications*, pages 680–690, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-29216-3.

[252] U. Raghavendra, U. Rajendra Acharya, Anjan Gudigar, Jen Hong Tan, Hamido Fujita, Yuki Hagiwara, Filippo Molinari, Pailin Kongmebhol, and Kwan Hoong Ng. Fusion of spatial gray level dependency and fractal texture features for the characterization of thyroid lesions. *Ultrasonics*, 77:110–120, 2017. ISSN 0041-624X. doi: https://doi.org/10.1016/j.ultras.2017.02.003.

[253] Reza Rahbari, Lisa Zhang, and Electron Kebebew. Thyroid cancer gender disparity. *Future Oncology*, 6(11):1771–1779, 2010. doi: 10.2217/fon.10.127.

[254] Andrik Rampun, Bryan W. Scotney, Philip J. Morrow, and Hui Wang. Breast mass classification in mammograms using ensemble convolutional neural networks. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6, 2018. doi: 10.1109/HealthCom.2018.8531154.

[255] Reese W. Randle, Norah M. Bushman, Jason Orne, Courtney J. Balentine, Elizabeth Wendt, Megan Saucke, Susan C. Pitt, Cameron L. Macdonald, Nadine P. Connor, and

Rebecca S. Sippel. Papillary thyroid cancer: The good and bad of the "good cancer". *Thyroid*, 27(7):902–907, 2017. doi: 10.1089/thy.2016.0632.

[256] Francesco Ranzato and Marco Zanella. Abstract interpretation of decision tree ensemble classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5478–5486, April 2020.

[257] Reza Rasti, Mohammad Teshnehlab, and Son Lam Phung. Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks. *Pattern Recognition*, 72: 381–390, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.08.004.

[258] Ruth Reátegui and Sylvie Ratté. Analysis of medical documents with text mining and association rule mining. In Álvaro Rocha, Carlos Ferrás, and Manolo Paredes, editors, *Information Technology and Systems*, pages 744–753, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11890-7.

[259] Adil Abdul Rehman and Khalid Alharthi. An introduction to research paradigms. *International Journal of Educational Investigations*, 3(8):51–59, 2016.

[260] Ying Ren, Yu He, and Linghua Cong. Application value of a deep convolutional neural network model for cytological assessment of thyroid nodules. *Journal of Healthcare Engineering*, 2021, 2021.

[261] Jia Rong. *Advanced pattern mining for complex data analysis*. PhD thesis, Deakin University, 2012.

[262] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[263] Patricia L. Rosenfield. The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science & Medicine*, 35(11): 1343–1357, 1992. ISSN 0277-9536. doi: https://doi.org/10.1016/0277-9536(92)90038-R. Special Issue Building Research Capacity for Health Social Sciences in Developing Countries.

[264] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, aug 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015720.

[265] E. M. Ruiz, T. Niu, M. Zerfaoui, M. Kunnimalaiyaan, P. L. Friedlander, A. B. Abdel-Mageed, and E. Kandil. A novel gene panel for prediction of lymph-node metastasis and recurrence in patients with thyroid cancer. *Surgery*, 167:73–79, 2020.

[266] Hajar Sadeghi, Mohammad Rafei, Masoud Bahrami, AliAkbar Haghdoost, and Yazdan Shabani. Attributable risk fraction of four lifestyle risk factors of thyroid cancer: a meta-analysis. *Journal of Public Health*, 40(2):e91–e98, 08 2017. ISSN 1741-3842. doi: 10.1093/pubmed/fdx088.

[267] Fatima M. Salman, Abu-Naser, and Samy S. Thyroid knowledge based system. *International Journal of Academic Engineering Research*, 3(5):11–20, 2019.

[268] Benedikt Schmidbauer, Karin Menhart, Dirk Hellwig, and Jirka Grosse. Differentiated thyroid cancer—treatment: state of the art. *International journal of molecular sciences*, 18(6):1292, 2017.

[269] Benedikt Schmidbauer, Karin Menhart, Dirk Hellwig, and Jirka Grosse. Differentiated thyroid cancer—treatment: State of the art. *International Journal of Molecular Sciences*, 18(6), 2017. ISSN 1422-0067. doi: 10.3390/ijms18061292.

[270] D. F. Schneider, H. Mazeh, H. Chen, and R. S. Sippel. Lymph node ratio predicts recurrence in papillary thyroid cancer. *The oncologist*, 18(2):157, 2013.

[271] Sara J Schonfeld, Choonsik Lee, and A Berrington De Gonzalez. Medical exposure to radiation and thyroid cancer. *Clinical Oncology*, 23(4):244–250, 2011.

[272] A. Schäffler. Hormone replacement after thyroid and parathyroid surgery. *Deutsches Ärzteblatt International*, 107(47), 2011.

[273] Martin G Seneviratne, Nigam H Shah, and Larry Chu. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, 6(2):45–47, 2020. ISSN 2055-8074. doi: 10.1136/bmjinnov-2019-000359.

[274] Atefeh Shahroudnejad, Roberto Vega, Amir Forouzandeh, Sharanya Balachandran, Jacob Jaremko, Michelle Noga, Abhilash Rakkunedeth Hareendranathan, Jeevesh Kapur, and Kumaradeven Punithakumar. Thyroid nodule segmentation and classification using deep convolutional neural network and rule-based classifiers. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3118–3121, 2021. doi: 10.1109/EMBC46164.2021.9629557.

[275] Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *International Conference on Machine Learning*, pages 2217–2225. PMLR, 2016.

[276] Xueda Shen, Xi Ouyang, Tianjiao Liu, and Dinggang Shen. Cascaded networks for thyroid nodule diagnosis from ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 145–154. Springer, 2020.

[277] Shyang-Rong Shih, Wei-Yih Chiu, Tien-Chun Chang, and Chin-Hsiao Tseng. Diabetes and thyroid cancer risk: literature review. *ExPerimental diabetes research*, 2012:1–7, 2012.

[278] Ilah Shin, Young Jae Kim, Kyunghwa Han, Eunjung Lee, Hye Jung Kim, Jung Hee Shin, Hee Jung Moon, Ji Hyun Youk, Kwang Gi Kim, , and Jin Young Kwak. Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland. *Ultrasonography*, 39(3):257, 2020.

[279] P. M. Silverman, G. E. Newman, M. Korobkin, J. B. Workman, A. V. Moore, and R. E. Coleman. Computed tomography in the evaluation of thyroid disease. *American journal of roentgenology*, 142(5):897–902, 1984.

[280] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[281] W.J. Simpson, S.E. McKinney, J.S. Carruthers, M.K. Gospodarowicz, S.B. Sutcliffe, and T. Panzarella. Papillary and follicular thyroid cancer: Prognostic factors in 1,578 patients. *The American Journal of Medicine*, 83(3):479–488, 1987. ISSN 0002-9343. doi: https://doi.org/10.1016/0002-9343(87)90758-3.

[282] Nikita Singh and Alka Jindal. A segmentation method and comparison of classification methods for thyroid ultrasound images. *International Journal of Computer Applications*, 50(11):43–49, 2012.

[283] American Cancer Society. Cancer facts & figures 2022, 2022. URL `https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics`.

[284] Ruoning Song, Long Zhang, Chuang Zhu, Jun Liu, Jie Yang, and Tong Zhang. Thyroid nodule ultrasound image classification through hybrid feature cropping network. *IEEE Access*, 8:64064–64074, 2020.

[285] Sebastian Stenman, Dmitrii Bychkov, Hakan Kücükel, Nina Linder, Caj Haglund, Johanna Arola, and Johan Lundin. Antibody supervised training of a deep learning based algorithm for leukocyte segmentation in papillary thyroid carcinoma. *IEEE Journal of Biomedical and Health Informatics*, 25(2):422–428, 2021. doi: 10.1109/JBHI.2020.2994970.

[286] Rose Stewart, Yit Jern Leang, Chhavi Raj Bhatt, Simon Grodski, Jonathan Serpell, and James C.Lee. Quantifying the differences in surgical management of patients with definitive and indeterminate thyroid nodule cytology. *European Journal of Surgical Oncology*, 46:252–257, 2020.

[287] George O Strawn. Scientific research: How many paradigms?. *Educause Review*, 47(3):26, 2012.

[288] Chrysostomos D. Stylios, Voula C. Georgopoulos, Georgia A. Malandraki, and Spyridoula Chouliara. Fuzzy cognitive map architectures for medical decision support systems. *Applied Soft Computing*, 8(3):1243–1251, 2008. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2007.02.022. Forging the Frontiers - - Soft Computing.

[289] Ming Sun, Qinglong Meng, Ting Wang, Tianci Liu, Ye Zhu, Jianfeng Qiu, and Weizhao Lu. Removal of manually induced artifacts in ultrasound images of thyroid nodules based on edge-connection and criminisi image restoration algorithm. *Computer Methods and Programs in Biomedicine*, 200:105868, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105868.

[290] K. S. Sundar, K. T. Rajamani, and S. S. S. Sai. Exploring image classification of thyroid ultrasound images using deep learning. In *International Conference on ISMAC in Computational Vision and Bio-Engineering*, pages 1635–1641. Springer, 2018.

[291] EINOSHIN SUZUKI. Undirected discovery of interesting exception rules. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(08):1065–1086, 2002.

[292] Keiji Suzuki, Vladimir Saenko, Shunichi Yamashita, and Norisato Mitsutake. Radiation-induced thyroid cancers: Overview of molecular signatures. *Cancers*, 11(9), 2019. ISSN 2072-6694. doi: 10.3390/cancers11091290.

[293] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition*, pages 1–9. IEEE, 2015.

[294] A Taccaliti, F Silvetti, G Palmonella, and M Boscaro. Genetic alterations in medullary thyroid cancer: diagnostic and prognostic markers. *Current genomics*, 12(8):618–625, 2011.

[295] Meera Tandan, Yogesh Acharya, Suresh Pokharel, and Mohan Timilsina. Discovering symptom patterns of covid-19 patients using association rule mining. *Computers in Biology and Medicine*, 131:104249, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104249.

[296] David Taniar, Wenny Rahayu, Vincent Lee, and Olena Daly. Exception rules in association rule mining. *Applied Mathematics and Computation*, 205(2):735–750, 2008. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2008.05.020. Special Issue on Advanced Intelligent Computing Theory and Methodology in Applied Mathematics and Computation.

[297] Yoshihiro Todoroki, Yutaro Iwamoto, Lanfen Lin, Hongjie Hu, and Yen-Wei Chen. Automatic detection of focal liver lesions in multi-phase ct images using a multi-channel amp; multi-scale cnn. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 872–875, 2019. doi: 10.1109/EMBC.2019.8857292.

[298] Jonathan Tritter. Mixed methods and multidisciplinary research in health care. *Researching health: Qualitative, quantitative and mixed methods*, pages 16–30, 2007.

[299] Chin-Hsiao Tseng. Thyroid cancer risk is not increased in diabetic patients. *PLOS ONE*, 7(12), 2012. doi: https://doi.org/10.1371/journal.pone.0053096.

[300] Vijay Vyas Vadhiraj, Andrew Simpkin, James O'Connell, Naykky Singh Ospina, Spyridoula Maraka, and Derek T. O'Keeffe. Ultrasound image classification of thyroid nodules using machine learning techniques. *Medicina*, 57(6), 2021. ISSN 1648-9144. doi: 10.3390/medicina57060527.

[301] Davy van de Sande, Michel E Van Genderen, Jim M. Smit, Joost Huiskens, Jacob J. Visser, Robert E. R. Veen, Edwin van Unen, Oliver Hilgers BA, Diederik Gommers, and Jasper van Bommel. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health & Care Informatics*, 29(1), 2022. doi: 10.1136/bmjhci-2021-100495.

[302] Douglas Van Nostrand. The benefits and risks of i-131 therapy in patients with well-differentiated thyroid cancer. *Thyroid*, 19(12):1381–1391, 2009. doi: 10.1089/thy.2009.1611.

[303] Carlo La Vecchia, Matteo Malvezzi, Cristina Bosetti, Werner Garavello, Paola Bertuccio, Fabio Levi, and Eva Negri. Thyroid cancer mortality and incidence: a global overview. *International journal of cancer*, 136(9):2187–2195, 2015.

[304] Durgadevi Velusamy and Karthikeyan Ramasamy. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Computer Methods and Programs in Biomedicine*, 198:105770, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105770.

[305] Durgadevi Velusamy and Karthikeyan Ramasamy. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. *Computer Methods and Programs in Biomedicine*, 198:105770, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105770.

[306] K Vondra, L Stárka, and R Hampl. Vitamin d and thyroid diseases. *Physiological research*, 64:S95, 2015.

[307] John P. Walsh. Managing thyroid disease in general practice. *Medical Journal of Australia*, 205(4):179–184, 2016.

[308] Hao Wang, Bin Song, Ningrong Ye, Jiliang Ren, Xilin Sun, Zedong Dai, Yuan Zhang, and Bihong T. Chen. Machine learning-based multiparametric mri radiomics for predicting the aggressiveness of papillary thyroid carcinoma. *European Journal of Radiology*, 122:108755, 2020. ISSN 0720-048X. doi: https://doi.org/10.1016/j.ejrad.2019.108755.

[309] L. Wang, Z. Q. Lin, and A. Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1): 1–12, 2020.

[310] Ping Wang, Long Lv, Feng Qi, and Feng Qiu. Increased risk of papillary thyroid cancer related to hormonal factors in women. *Tumor Biology*, 36(7):5127–5132, 2015.

[311] Tracy S. Wang, Kevin Cheung, Forough Farrokhyar, Sanziana A. Roman, and Julie Ann Sosa. A meta-analysis of the effect of prophylactic central compartment neck dissection on locoregional recurrence rates in patients with papillary thyroid cancer. *Annals of surgical oncology*, 20(11):3477–3483, 2013.

[312] Xuanqi Wang, Bin Cao, Dejian Wei, Jing Liu, and Hui Cao. Diagnosis of thyroid nodules based on lightweight residual network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3875–3881, 2021. doi: 10.1109/BIBM52615.2021.9669574.

[313] Yunjun Wang, Qing Guan, Iweng Lao, Li Wang, Yi Wu, Duanshu Li, Qinghai Ji, Yu Wang, Yongxue Zhu, Hongtao Lu, and Jun Xiang. Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. *Annals of translational medicine*, 7(18), 2019.

[314] Mary H. Warda, Briseis A. Kilfoya, Peter J. Weyerb, Kristin E. Andersonc, Aaron R.

Folsomc, and James R. Cerhan. Nitrate intake and the risk of thyroid cancer and thyroid disease. *Epidemiology*, 21:255–264, 2010. doi: 10.1097/EDE.0b013e3181d6201d.

[315] Leonard Wartofsky. Increasing world incidence of thyroid cancer: increased detection or higher radiation exposure? *Hormones*, 9(2):103–108, 2010.

[316] Ran Wei, Hao Wang, Lanyun Wang, Wenjuan Hu, Xilin Sun, Zedong Dai, Jie Zhu, Hong Li, Yaqiong Ge, and Bin Song. Radiomics based on multiparametric mri for extrathyroidal extension feature prediction in papillary thyroid cancer. *BMC Med Imaging*, 21(20), 2021. doi: https://doi.org/10.1186/s12880-021-00553-z.

[317] Marjorie Cecilia Weiss. Diagnostic decision making: The last refuge for general practitioners? *Social Science & Medicine*, 73(3):375–382, 2011. ISSN 0277-9536. doi: https://doi.org/10.1016/j.socscimed.2011.05.038.

[318] Kirsty Williamson and Graeme Johanson. *Research methods: Information, systems, and contexts*. Chandos Publishing, 2017.

[319] Joseph J. Wiltshire, Thomas M. Drake, Lesley Uttley, and Sabapathy P. Balasubramanian. Systematic review of trends in the incidence rates of thyroid cancer. *Thyroid*, 26(11):1541–1552, 2016. doi: 10.1089/thy.2016.0100.

[320] Andrea Winquist and Kyle Steenland. Perfluorooctanoic acid exposure and thyroid disease in community and worker cohorts. *Epidemiology*, 25(2):255–264, 2014. ISSN 10443983.

[321] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

[322] Hongxun Wu, Zhaohong Deng, Bingjie Zhang, Qianyun Liu, and Junyong Chen. Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography. *American Journal of Roentgenology*, 207(4): 859–864, 2016.

[323] Xindong Wu, Chengqi Zhang, and Shichao Zhang. Efficient mining of both positive and negative association rules. *ACM Trans. Inf. Syst.*, 22(3):381–405, jul 2004. ISSN 1046-8188. doi: 10.1145/1010614.1010616.

[324] Yao-yuan Wu, Chao Wei, Chuan-bin Wang, Nai-yu Li, Ping Zhang, and Jiang-ning Dong. Preoperative prediction of cervical nodal metastasis in papillary thyroid carcinoma: Value of quantitative dual-energy ct parameters and qualitative morphologic features. *American Journal of Roentgenology*, 216(5):1335–1343, 2021. doi: 10.2214/AJR.20.23516.

[325] Constance Xhaard, Yan Ren, Enora Clero, Stephane Maillard, Pauline Brindel, Frederique Rachedi, Frederique Bost-Bezeaud, Francoise Doyon, Carole Rubino, and Florent de Vathaire. Differentiated thyroid carcinoma risk factors in french polynesia. *Asian Pacific journal of cancer prevention*, 15(6):2675–2680, 2014.

[326] Constance Xhaard, Florent de Vathaire, Enora Cléro, Stéphane Maillard, Yan Ren, Françoise Borson-Chazot, Geneviève Sassolas, Claire Schvartz, Marc Colonna, Brigitte Lacour, Arlette Danzon, Michel Velten, Emilie Marrer, Laurent Bailly, Eugènia Mariné Barjoan, Martin Schlumberger, Jacques Orgiazzi, Elisabeth Adjadj, and Carole Rubino. Anthropometric risk factors for differentiated thyroid cancer in young men and women from eastern france: A case-control study. *American Journal of Epidemiology*, 182(3): 202–214, 06 2015. ISSN 0002-9262. doi: 10.1093/aje/kwv048.

[327] Jianfu Xia, Huiling Chen, Qiang Li, Minda Zhou, Limin Chen, Zhennao Cai, Yang Fang, and Hong Zhou. Ultrasound-based differentiation of malignant and benign thyroid nodules: An extreme learning machine approach. *Computer Methods and Programs in Biomedicine*, 147:37–49, 2017. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2017.06.005.

[328] Kele Xu, Dawei Feng, Haibo Mi, Boqing Zhu, Dezhi Wang, Lilun Zhang, Hengxing Cai, and Shuwen Liu. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In *Pacific Rim conference on multimedia*, pages 14–23. Springer, 2018.

[329] Karkoubi Y., Moradi Gh., Sharifi P., and Ghafouri Sh. Assessment of thyroid cancer risk factors in kurdistan province. *Scientific Journal of Kurdistan University of Medical Sciences*, 23(3):10–18, 2018.

[330] Qingsen Yan, Bo Wang, Wei Zhang, Chuan Luo, Wei Xu, Zhengqing Xu, Yanning Zhang, Qinfeng Shi, Liang Zhang, and Zheng You. Attention-guided deep neural network with

multi-scale feature fusion for liver vessel segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2629–2642, 2021. doi: 10.1109/JBHI.2020.3042069.

[331] Charles Q. Yang, Lauren Gardiner, Huan Wang, Matthew T. Hueman, and Dechang Chen. Creating prognostic systems for well-differentiated thyroid cancer using machine learning. *Frontiers in Endocrinology*, 10:288, 2019. ISSN 1664-2392. doi: 10.3389/fendo.2019.00288.

[332] Qinghan Yang, Chong Geng, Ruyue Chen, Chen Pang, Run Han, Lei Lyu, and Yuang Zhang. Dmu-net: Dual-route mirroring u-net with mutual learning for malignant thyroid nodule segmentation. *Biomedical Signal Processing and Control*, 77:103805, 2022. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2022.103805.

[333] Xi Yang, Shupeng Qiu, and Qiong Luo. Feature-based discrimination of thyroid cancer on ultrasound images. In *2020 IEEE 3rd International Conference on Electronics Technology (ICET)*, pages 834–839, 2020. doi: 10.1109/ICET49382.2020.9119560.

[334] Yohwan Yeo, Seung-Hyun Ma, Yunji Hwang, Pamela L. Horn-Ross, Ann Hsing, Kyu-Eun Lee, Young Joo Park, Do-Joon Park, Keun-Young Yoo, and Sue K. Park. Diabetes mellitus and risk of thyroid cancer: A meta-analysis. *PLOS ONE*, 9(6):1–11, 06 2014. doi: 10.1371/journal.pone.0098135.

[335] Y. J. Yoo, E. J. Ha, Y. J. Cho, H. L. Kim, M. Han, and S. Y. Kang. Computer-aided diagnosis of thyroid nodules via ultrasonography: initial clinical experience. *Korean journal of radiology*, 19(4):665, 2018.

[336] Inyoung Youn, Eunjung Lee, Jung Hyun Yoon, Hye Sun Lee, Mi-Ri Kwon, Juhee Moon, Sunyoung Kang, Seul Ki Kwon, Kyong Yeun Jung, Young Joo Park, et al. Diagnosing thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance cytology with the deep convolutional neural network. *Scientific Reports*, 11 (1):1–9, 2021.

[337] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, May 2016.

[338] Xianglong Zeng, Haiquan Chen, Yuan Luo, and Wenbin Ye. Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network. *IEEE Access*, 7:30744–30753, 2019. doi: 10.1109/ACCESS.2019.2903171.

[339] Tarek Zetoune, Xavier Keutgen, Daniel Buitrago, Hasan Aldailami, Huibo Shao, Madhu Mazumdar, Thomas J. Fahey III, and Rasa Zarnegar. Prophylactic central neck dissection and local recurrence in papillary thyroid cancer: a meta-analysis. *Annals of surgical oncology*, 17(12):3287–3293, 2010.

[340] Mimi Zhai, Dan Zhang, Jianhai Long, Yi Gong, Fei Ye, Sushun Liu, and Yamin Li. The global burden of thyroid cancer and its attributable risk factor in 195 countries and territories: A systematic analysis for the global burden of disease study. *Cancer Medicine*, 10 (13):4542–4554, 2021. doi: https://doi.org/10.1002/cam4.3970.

[341] Bin Zhang, Jie Tian, Shufang Pei, Yubing Chen, Xin He, Yuhao Dong, Lu Zhang, Xiaokai Mo, Wenhui Huang, Shuzhen Cong, and Shuixing Zhang. Machine learning–assisted system for thyroid nodule diagnosis. *Thyroid*, 29(6):858–867, 2019.

[342] Rong Zhang, Qiufang Liu, Hui Cui, Xiuying Wang, Shaoli Song, Gang Huang, and Dagan Feng. Thyroid classification via new multi-channel feature association and learning from multi-modality mri images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 277–280, 2018. doi: 10.1109/ISBI.2018.8363573.

[343] Yiwen Zhang, Haoran Lai, and Wei Yang. Cascade unet and ch-unet for thyroid nodule segmentation and benign and malignant classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–134. Springer, 2020.

[344] Hengqiang Zhao, Hehe Li, and Tao Huang. High iodine intake and central lymph node metastasis risk of papillary thyroid cancer. *Journal of Trace Elements in Medicine and Biology*, 53:16–21, 2019. ISSN 0946-672X. doi: https://doi.org/10.1016/j.jtemb.2019.01.015.

[345] Hongbo Zhao, Chang Liu, Jing Ye, Lufan Chang, Qing Xu, Bowen Shi, Lulu Liu, Yili Yin, and Binbin Shi. A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on ct images. *Endokrynologia Polska*, 72(3):217 – 225, 2021. ISSN 2299-8306. doi: 10.5603/EP.a2021.0015.

[346] Junyu Zhao, Haipeng Wang, Zhongwen Zhang, Xiaojun Zhou, Jinming Yao, Rui Zhang, Lin Liao, and Jianjun Dong. Vitamin d deficiency as a risk factor for thyroid cancer: A meta-analysis of case-control studies. *Nutrition*, 57:5–11, 2019. ISSN 0899-9007. doi: https://doi.org/10.1016/j.nut.2018.04.015.

[347] Sitong Zhao, Xiaomeng Jia, Xiaojing Fan, Ling Zhao, Ping Pang, Yajing Wang, Yukun Luo, Fulin Wang, Guoqing Yang, Xianling Wang, Weijun Gu, Li Zang, Yu Pei, Jin Du, Jianming Ba, Jingtao Dou, Yiming Mu, and Zhaohui Lyu. Association of obesity with the clinicopathological features of thyroid cancer in a large, operative population: a retrospective case-control study. *Medicine*, 98(50):e18213, 2019. doi: 10.1097/MD.0000000000018213.

[348] Zijian Zhao, Congmin Yang, Qian Wang, Huawei Zhang, Linlin Shi, and Zhiwen Zhang. A deep learning-based method for detecting and classifying the ultrasound images of suspicious thyroid nodules. *Medical Physics*, 48(12):7959–7970, 2021.

[349] Zuopeng Zhao, Chen Ye, Yanjun Hu, Ceng Li, and Xiaofeng Li. Cascade and fusion of multitask convolutional neural networks for detection of thyroid nodules in contrast-enhanced ct. *Computational intelligence and neuroscience*, 2019.

[350] Hao Zheng, Yizhe Zhang, Lin Yang, Peixian Liang, Zhuo Zhao, Chaoli Wang, and Danny Z. Chen. A new ensemble learning framework for 3d biomedical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5909–5916, July 2019.

[351] Shaohua Zheng, Zhiqiang Shen, Chenhao Pei, Wangbin Ding, Haojin Lin, Jiepeng Zheng, Lin Pan, Bin Zheng, and Liqin Huang. Interpretative computer-aided lung cancer diagnosis: From radiology analysis to malignancy evaluation. *Computer Methods and Programs in*

Biomedicine, 210:106363, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106363.

[352] Xiangyang Xue Zhiqiang Shen, Zhankui He. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4886–4893, July 2019.

[353] Chenghao Zhou, Ji Geng, and Hengyu Li. Segmentation of ultrasound thyroid nodules based on encoder-decoder structure. In *2021 International Conference on Internet, Education and Information Technology (IEIT)*, pages 333–337, 2021. doi: 10.1109/IEIT53597.2021.00079.

[354] Shujun Zhou, Hong Wu, Jie Gong, Ting Le, Hao Wu, Qin Chen, and Zenglin Xu. Mark-guided segmentation of ultrasonic thyroid nodules using deep learning. In *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine*, ISICDM 2018, page 21–26. Association for Computing Machinery, 2018. ISBN 9781450365338. doi: 10.1145/3285996.3286001.

[355] Yan Zhou, Guo-Yi Su, Hao Hu, Ying-Qian Ge, Yan Si, Mei-Ping Shen, Xiao-Quan Xu, and Fei-Yun Wu. Radiomics analysis of dual-energy ct-derived iodine maps for diagnosing metastatic cervical lymph nodes in patients with papillary thyroid cancer. *European radiology*, 30(11):6251–6262, 2020.

[356] Zhihua Zhou. *Ensemble Learning*, pages 181–210. Springer Singapore, Singapore, 2021. ISBN 978-981-15-1967-3. doi: 10.1007/978-981-15-1967-3_8.

[357] Lu-Cheng Zhu, Yun-Liang Ye, Wen-Hua Luo, Meng Su, Hang-Ping Wei, Xue-Bang Zhang, Juan Wei, and Chang-Lin Zou. A model to discriminate malignant from benign thyroid nodules using artificial neural network. *PLoS One*, 8(12):e82211, 2013.

[358] Y. Zhu, Z. Fu, and J Fei. An image augmentation method using convolutional network for thyroid nodule classification by transfer learning. In *IEEE International Conference on Computer and Communications (ICCC)*, pages 1819–1823. IEEE, 2017.

[359] Yi-Cheng Zhu, Alaa AlZoubi, Sabah Jassim, Quan Jiang, Yuan Zhang, Yong-Bing Wang, Xian-De Ye, and Hongbo DU. A generic deep learning framework to classify thyroid and breast lesions in ultrasound images. *Ultrasonics*, 110:106300, 2021. ISSN 0041-624X. doi: https://doi.org/10.1016/j.ultras.2020.106300.

[360] Michael B. Zimmermann and Valeria Galetti. Iodine intake as a risk factor for thyroid cancer: a comprehensive review of animal and human studies. *Thyroid research*, 8(1):1–21, 2015. doi: https://doi.org/10.1186/s13044-015-0020-8.

[361] Vladan Zivaljevic, Nikola Slijepcevic, Sandra Sipetic, Ivan Paunovic, Aleksandar Diklic, Goran Zoric, and Nevena Kalezic. Risk factors for well-differentiated thyroid cancer in men. *Tumori Journal*, 99(4):458–462, 2013. doi: 10.1177/030089161309900403.

[362] J. Zuluaga-Gomez, Z. Al Masry, K. Benaggoune, S. Meraghni, and N. Zerhouni. A cnn-based methodology for breast cancer diagnosis using thermal images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–15, 2020.

[363] Şaban Öztürk, Umut Özkaya, Bayram Akdemir, and Levent Seyfi. Convolution kernel size effect on convolutional neural network in histopathological image processing applications. In *2018 International Symposium on Fundamentals of Electrical Engineering (ISFEE)*, pages 1–5, 2018. doi: 10.1109/ISFEE.2018.8742484.

Monash University Human Research Ethics Committee

Approval Certificate

This is to certify that the project below was considered by the Monash University Human Research Ethics Committee. The Committee was satisfied that the proposal meets the requirements of the *National Statement on Ethical Conduct in Human Research* and has granted approval.

Project ID:	24704
Project Title:	Thyroid Cancer Prediction
Chief Investigator:	Assoc Professor Vincent Lee
Approval Date:	27/07/2020
Expiry Date:	27/07/2025

Terms of approval - failure to comply with the terms below is in breach of your approval and the *Australian Code for the Responsible Conduct of Research*.

1. The Chief Investigator is responsible for ensuring that permission letters are obtained, if relevant, before any data collection can occur at the specified organisation.
2. Approval is only valid whilst you hold a position at Monash University.
3. It is responsibility of the Chief Investigator to ensure that all investigators are aware of the terms of approval and to ensure the project is conducted as approved by MUHREC.
4. You should notify MUHREC immediately of any serious or unexpected adverse effects on participants or unforeseen events affecting the ethical acceptability of the project.
5. The Explanatory Statement must be on Monash letterhead and the Monash University complaints clause must include your project number.
6. Amendments to approved projects including changes to personnel must not commence without written approval from MUHREC.
7. Annual Report - continued approval of this project is dependent on the submission of an Annual Report.
8. Final Report - should be provided at the conclusion of the project. MUHREC should be notified if the project is discontinued before the expected completion date.
9. Monitoring - project may be subject to an audit or any other form of monitoring by MUHREC at any time.
10. Retention and storage of data - The Chief Investigator is responsible for the storage and retention of the original data pertaining to the project for a minimum period of five years.

Kind Regards,

Professor Nip Thomson

Chair, MUHREC

CC: Ms Xinyu Zhang, Dr Jackie Rong, Dr Feng Liu

List of approved documents:

Document Type	File Name	Date	Version
Supporting Documentation	Permission Letter	21/06/2020	1

Source Code

This thesis ensures the reproducibility, and all the code and partial datasets are available on GitHub:

1. Association Rule Mining
    (a) Exception Rules Extraction
    (b) Faster Apriori
    (c) Terminology Extraction
    (d) CN Dataset - Blood
2. Medical Image Segmentation
3. Medical Image Augmentation
4. CAD Implementation
    (a) Binary Classification
    (b) Multi-class Classification
5. Multi-channel CNN Architectures
    (a) SIDC
    (b) DIDC
    (c) Four-channel
    (d) Multi-channel Multi-class Classification
6. Unified Model Selection
    (a) Individual Leaner Selection
    (b) Weighted Ensemble Averaging Model