Vision Transformer-based Context-Aware System for Lingual Ultrasound in Digital Health Ecosystem

by

Khalid Al-hammuri
B.Sc., Yarmouk University, 2012
M.Sc., University of Victoria, 2019

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

Vision Transformer-based Context-Aware System for Lingual Ultrasound in Digital Health Ecosystem

by

Khalid Al-hammuri
B.Sc., Yarmouk University, 2012
M.Sc., University of Victoria, 2019

Supervisory Committee

Dr. Fayez Gebali, Co-Supervisor
(Electrical and Computer Engineering, University of Victoria)

Dr. Awos Kanan, Co-Supervisor
(Electrical and Computer Engineering, Prince Sumaya University for Technology)

Dr. Afzal Suleman, Outside Member
(Mechanical Engineering Department, University of Victoria)

Supervisory Committee

Dr. Fayez Gebali, Co-Supervisor
(Electrical and Computer Engineering, University of Victoria)

Dr. Awos Kanan, Co-Supervisor
(Electrical and Computer Engineering, Prince Sumaya University for Technology)

Dr. Afzal Suleman, Outside Member
(Mechanical Engineering Department, University of Victoria)

ABSTRACT

The complex nature of modern healthcare systems and the widespread distribution of healthcare infrastructure made the interoperability within healthcare information system challenging. This could poses security risks, missing data, miscommunication , in addition to the human and technical-based errors. This dissertation focuses on utilizing an advanced AI system to overcome the challenges of clinical analysis, data confidentiality, availability, and integrity. There are three main contributions of this research. First, implement TongueTransUNet, which is a well-managed architecture that utilizes a vision transformer, UNet encoder-decoder convolutional neural network, contrastive loss and quality control process supported with human-reinforcement feedback to extract tongue fingerprint. Second, design ZTCloudGuard for access control within the telehealth cloud-based eco-system between. The architecture manage users, devices, and output attributes by deriving a score to assess the mutual relationship considering semantic and syntactic analysis. Third,utilize hybrid qualitative and quantitative evaluation metrics and conduct comparative analysis to other related research. The main applications to this research are minimizing medical errors, protecting healthcare practitioners, detecting unrelated input and undesired output. An ablation study using synthetic healthcare information attributes and word2vec model was conducted to judge the

model results. The outcomes showed robustness and enhancement by focusing on high-quality input and rejecting unacceptable data. If the automatic process fails or goes below a predefined threshold, an extra reinforcement verification layer is introduced to the algorithm to add manual and human feedback.

Contents

List of Tables

List of Figures

Abbreviations

2D:	Two dimensions
3D:	Three dimensions
AI:	Artificial intelligence
ASM:	Active shape model
BERT:	Bidirectional encoder representations from transformers
CNN:	Convolutional neural networks
CT:	Computed tomography
CL:	Contrastive learning
CRC:	Colorectal Cancer
CW:	Complex wavelet
DL:	Deep learning
EV :	Evaluation
EEG:	Electroencephalography
EMG:	Electromyography
EPG:	Electropalatography
EMA:	Electromagnetic articulatory
ECG:	Electrocardiography
FCN:	Fully convolutional network
GT:	Ground truth
GAN:	Generative adversarial network
HIS:	Hospital information system
IoT:	Internet of things
K:	Key
LSTM:	Long short-term memory
MLP:	Multi-layer perceptron
ML:	Machine learning
MSA:	Multi-head self attention

Continue for Abbreviation

MRI:	Magnetic resonance imaging
MSD:	Mean sum of distances
MSE:	Mean squared error
MRI:	Magnetic resonance imaging
NLP:	Natural language processing
Norm:	Normalization layer
PCA:	Principal component analysis
Q:	Query
RIS:	Radiology information system
RL:	Reinforcement learning
RMSE:	Root-mean-square error
SSIM:	Structural similarity index measure
SSI:	Silent-speech interface
US:	Ultrasound
V:	Value
ViT:	Vision Transformer

Chapter 1

Introduction and Research Statement

1.1 Introduction and problem statement

The emerging need for smart healthcare systems to adapt to the fast-growing demands on resources and integrate the healthcare facilitates to the new advanced technology creates new challenges that need further research and development. The main obstacles are utilizing telehealth services and their associated security for the distributed endpoints to protect patient information privacy and maintain service quality. The solutions and remedies can be achieved using current technological advancements in areas like the Cloud, AI, 5G networks, Blockchain, and distributed IoT devices.
The dissertation focuses on solving three main challenges within the distributed cloud AI healthcare ecosystem. Healthcare data clinical context analysis, data confidentiality, availability and integrity are the main problems the dissertation addresses.

Challenge one (Clinical Images Analysis): The main purpose of the healthcare practitioner is to have advanced tools to analyze patient clinical data in a fast, accurate and useful manner. This is difficult to achieve due to the variation in human nature and data complexity alongside the limitation of medical devices that may produce noisy data. To solve this challenge, the research focuses on analyzing lingual ultrasound images as a case study. The dissertation presents TongueTransUNet, which is a quality control and Vision Transformer (ViT)-based context-aware system used to analyze lingual ultrasound images for clinical image

segmentation. TongueTransUNet also supported by the human-based reinforcement feedback. The proposed architecture is used to extract tongue contour from complex and noisy ultrasound images. The extracted features serve as a unique signature that serve as a tongue fingerprint that can help to identify subject and for language and lingual behavioural understanding.

Challenge two (Data Confidentiality and Availability): The next stage after acquiring and analyzing the healthcare information is to send them to the final destination, either a user, device, storage or another computing facility for further analysis. Data confidentiality is important to guarantee the data is accessed only by authorized users or devices. On the other hand, data availability ensures that the data is always available for the designated users when needed. To handle the second challenge of securing data confidentiality and availability, the research proposes a secure zero-trust context-aware healthcare access management protocol for ultrasound images and other IoT devices within the cloud-based ecosystem. The proposed zero trust context-aware framework is unique and handles the complicated nature of communication between different medical devices and users that are either robots or humans within the cloud-based healthcare infrastructure by limiting the data access to authorized users and resources only.

Challenge three (Data Integrity): Transmitting medical data to distributed devices in a cloud-based system located in different geographical locations has some limitations. The data could suffer from network data loss, connection interruption, man-in-the-middle cyber attacks and other malicious attacks that may pose wrong information or degrade the data quality. The research addresses data integrity by proposing unique qualitative and quantitative evaluation measures to validate the medical tongue image segmentation. The evaluation measures ensure that the transmitted health data are not tampered with or altered during the data journey between users, IoT devices and cloud resources.

The main dissertation research application of using ViT in digital health is for tongue segmentation and telehealth security. The dissertation will evaluate and compare the usability of ViT to extract the tongue from the complex oropharyngeal structure to avoid fast movement, noise and missing data. The extracted contour will be used to further analyze medical ultrasound images and telehealth communication protocols. The context of the extracted images will be used for access control management of the IoT medical devices, users and data output during data exchange between different resources like the electronic health records, endpoint

devices and picture archiving system.

The application of tongue contour segmentation is also useful and applicable for different applications. The segmented contour can be used for tongue tracking to produce a language model from studying speech behaviour. It also can be used for second language learning, IoT device recognition, speech disorder analysis, and neural-related diseases like Parkinson's and dementia.

This type of application and analysis will support the design of the smart hospital health information system. Smart hospitals are emerging as an essential demand for modern healthcare infrastructure. Having smart hospitals also requires using smart medical devices that can communicate through the digital health communication system. Medical imaging occupies 90% of medical data. The main applications are minimizing medical errors, protecting healthcare practitioners, detecting unrelated input and undesired output.

1.2 Research Outcomes and Achievements

To overcome the challenges in Sec. 1.1, the research achievements and contributions are highlighted in three main points.

Contributions One: Construct TongueTransUNet which is a well-manged architecture that utilizes ViT, UNet, contrastive loss and other quality control process to enhance the performance of model data input and output. The architecture also utilizes human-reinforcement feedback to further enhance the results. The details explained in Chapters 2, 3, 4.

Contributions Two: Design ZTCloudGuard which is a zero-trust context aware access management framework. It help to verify confidentiality, availability and integrity of the data interoperability between users, devices and data output within distributed telehealth system. ZTCloudGuard utilize scoring system to prevent or alleviate medical errors using semantic and syntactic analysis. ZTCloudGuard design explained in Chapter 5.

Contributions Three: Propose hybrid qualitative and quantitative approach to judge on the final data output. In addition the research conducts experimental and comparative analysis to compare the proposed with to other related researches in the

field. This is important to ensure transmitting accurate data without loss of essential health information. This contribution detailed in Chapter 6.

1.3 Summary of completed work

This section presents the published work toward achieving the project goals. Each point contains details of the published article in the following itemized points.

1. **Article title:** Al-Hammuri K, Gebali F, Thirumarai Chelvan I, Kanan A. Tongue contour tracking and segmentation in lingual ultrasound for speech recognition: A review. Diagnostics. 2022 Nov 15;12(11):2811; [10]. [Link].
   Publisher: Diagnostics 2022, 12(11), 2811;
   Status: Published;
   Publication date: November 15, 2022.

2. **Article title:** Al-Hammuri K, Gebali F, Kanan A, Chelvan IT. Vision transformer architecture and applications in digital health: a tutorial and survey. Visual computing for industry, biomedicine, and art. 2023 Jul 10;6(1):14; [11]. [Link]
   Publisher: Springer (Visual Computing for Industry, Biomedicine and Arts);
   Status: Published;
   Publication date: July 10, 2023;

3. **Article title:** Khalid Al-hammuri, Fayez Gebali, Awos Kanan et al. Tongue-TransUNet: Toward Effective Tongue Contour Segmentation Using well-managed Dataset.
   Status: Accepted;

4. **Article title:** Al-Hammuri K, Gebali F, Kanan A. ZTCloudGuard: Zero Trust Context-Aware Access Management Framework to Avoid Medical Errors in the Era of Generative AI and Cloud-Based Health Information Ecosystems. AI. 2024 Jul 8;5(3):1111-31; [12]. [Link]
   Publisher: MDPI AI;
   Status: Published;
   Publication date: July 8, 2024;

1.4 Dissertation outline and structure

This section is a brief summary that outlines the rest of the dissertation chapters in itemized points to describe the PhD research.

Chapter 2 Background of the work, which includes an overview and literature review of the tongue contour tracking techniques.

Chapter 3 Discuss the Vision-Transformer architecture design and its applications in digital health applications.

Chapter 4 Proposes methodology using TongueTransUNet architecture of the tongue contour segmentation and features extraction from ultrasound images using well-managed quality control process. This chapter also includes information about data sets and data acquisition.

Chapter 5 This chapter presents the ZTCloudGuard design which is based on the zero trust context-aware system. The system used to manage the communication between users, devices and output to manage medical device access within the cloud-based healthcare infrastructure.

Chapter 6 The chapter include evaluation and comparative analysis to other techniques. The chapter utilizes hybrid quantitative and qualitative analysis of different tongue contour segmentation algorithms.

Chapter 2

Background of the Tongue Contour Segmentation in Lingual Ultrasound

This chapter highlights the main bio-sensors used for tongue contour analysis. The chapter also discusses the two main techniques for tongue contour segmentation. Traditional computer vision and machine learning-based techniques are the two main methodologies harnessed in this research.

2.1 Overview of different biosensors used for speech analysis

Studying tongue movement during speech is essential to the understanding of human articulation. Different approaches are used to study speech; some rely on a single sensor [13, 14, 15, 8, 16], and others use hybrid techniques [17, 18, 19]. Due to medical imaging modalities advancement and impressive capabilities, linguistic researchers rely on the medical ultrasound system to capture tongue motion during speech [20]. Ultrasound imaging is considered the most efficient methodology in terms of safety and portability.

However, magnetic resonance imaging (MRI) has a better resolution and can provide more information about the soft tissues [21], vocal tract, and craniofacial structure [22, 23]. MRI is used for real-time image acquisition [21, 24, 25] to visualize the vocal tract either in 2D or 3D orientation [26, 27] and enhance the speech analysis. However, MRI is huge in size and very expensive compared to ultrasound. It requires a special arrangement and a long scanning time, making it im-

practical for most of the day-to-day uses of speech analysis to limit its application for particular research or clinical studies.

On the other hand, X-ray [28, 29, 30] and CT [31, 32, 33, 34, 35] systems are cheaper than MRI, they have a reasonable resolution, and they have many applications as well. X-ray is used for tongue contour extraction [36, 37]; it is also used for tongue contour image synthesis to create articulation copy [38] or combining physiological models to fit X-ray images. An X-ray system is also beneficial for capturing images of the whole vocal tract [39] and nonrigid articulatory structures [40]. CT scan has a wide variety of applications compared to conventional X-rays.

CT scan is used in clinical studies of oral-cavity-related disorders such as sleep apnea [41, 32]. CT images are also used to estimate the tongue volume within the oral cavity [42, 43, 44, 45]. Furthermore, CT is applied in advanced surgical procedures as it is beneficial for image registration [46]; augmented reality and CT images are also combined to guide transoral robotic surgery [47]. In addition, CT-mapped 3D images of different tongue types have been used in clinical applications of tongue cancers [48]. However, CT and X-rays are larger in size compared to ultrasound, and they have a radiation danger which requires a strictly yearly radiation dose limit to prevent harmful radiation for humans. At the same time, ultrasound is safer and has no radiation danger to the user.

In addition to the medical imaging systems, biosignal sensors are also utilized for speech analysis and related studies. Types of biosignal sensors [18] are electromagnetic articulatory (EMA), permanent magnet articulography (PMA), electropalatography (EPG), electromyography (EMG), electroencephalography (EEG).

EMA [49, 50, 51] is useful to localize the movement within the vocal tract by using electromagnetic transmitter coils to track the position of the attached electromagnetic sensors on the tongue, lips, and jaw. EMA may provide either a 2D or 3D landmark localization in milliseconds, but the system operation is complex and uncomfortable to be used in all cases on a daily basis; it might be more usable for conducting clinical studies at research centres. In ultrasound research, EMA data are used to build a prediction machine learning model to guide ultrasound tracking to minimize the effect of missing data.

On the other hand, PMA [52, 53, 54, 55, 56] is a technique to capture articulator displacement by using a permanent magnet on the tongue and detecting the magnetic signal using a wearable sensor. It is useful for speech recognition tasks, and

the reported word detection accuracy is around 90% [52]. Unlike EMA, PMA does not have wires and has a reverse transmitter-receiver arrangement to make it more convenient [18]. However, PMA sensor configuration is not convenient, and it is difficult to maintain the same position reference for all cases.

EPG is used for tongue tracking and speech therapy [18, 57]. Moreover, EPG information is also applied to get an accurate image registration by a CT scanner [45]. Furthermore, EPG can be combined with audio signals for speech generation and speech enhancement applications [58]. EPG uses a hard plate beneath the tongue to detect the contact between the tongue and the array of sensors in the plate. The hard plate requires a specialized dentist to get a measure, as it should be custom-designed for each patient. However, EPG can give some information about the tongue motion, but it is not practical, and limited data can be acquired from it compared to ultrasound.

EMG for speech recognition [59, 60, 61] is more convenient and safer than EMA, PMA and EPG as it uses surface electrodes on the face without any invasive measures. EMG is a system that detects the muscles' electric activity and its nerves' biosignals [62]. The detected signals can give an indication of the muscles' health [63]. However, in the case of speech recognition, the muscles' movement can indicate the speech behaviour and its relationship with the tongue muscle motion [64]. Moreover, EMG can be used to translate hand gestures for a speech to help people with speech impediments [65].

Studying brain electrical activity using EEG is useful for speech analysis. The acoustic sound stimulates the auditory cortex in the brain, which generates electrical signals that can be detected by the electrodes or small metal plates attached to the scalp. Different research studies have proposed to analyze EEG signals and extract the relationship between brain signals and speech behaviour [66, 67, 68, 69]. Although EEG can provide information about speech patterns, the nature of the EEG signal is complex and susceptible to noise, which makes the part of the EEG complex signal relating to the auditory system difficult to separate from other electrical activities of the brain [18, 70]. Many advanced techniques have been proposed to alleviate this issue by proposing artifact removal [71, 72] or incorporating advanced deep learning techniques such as a Transformer model and a generative adversarial network analysis [73, 74].

2.2 Ultrasound imaging in speech recognition fundamentals

An ultrasound system is portable, safe, and convenient, making it efficient for real-time image acquisition inside or outside hospitals. Researchers and clinical linguists have widely adopted the use of lingual ultrasound for different applications. Some of these applications include using it as visual feedback for second language teaching [75, 76], speech remediation to correct articulation for people with speech disabilities [77], speech-related disorders such as autism [15, 78, 79], articulation research and analysis [20, 75, 80], swallowing studies [81], tongue 3D modelling [82], and silent speech interface [83, 84, 85, 17].

Furthermore, ultrasound imaging analysis is used in many applications in medical imaging analysis for object detection and segmentation. Some of these applications are in the field of cardiology, in which researchers obtain echocardiography images for the heart to help cardiologists identify the health status of the heart [86, 87]. Echocardiography image segmentation is beneficial for measuring the left ventricle volume and estimating its blood ejection fraction. It is also useful for examining heart valve performance. Moreover, ultrasound is also one of the safest and most efficient tools for studying breast cancer and assisting with cancer biopsy.

Ultrasound images could help physicians examine breast tissues to identify if a cancerous mass is benign or malignant, either in two-dimension (2D) images [86, 88, 89] or three-dimension (3D) images [90]. A portable ultrasound system is also used in healthcare facilities to assist in intravascular procedures [91, 92]. Obstetrics and gynecology use ultrasound systems on a daily basis to examine and mentor pregnant women's health and fetus growth [93, 94]. Furthermore, ultrasound is also used to detect ovarian tumours, which is one of the main diseases that affect women's health [95].

Figure 2.1 visualizes the placement of the ultrasonic transducer beneath the chin and the propagation of the acoustic wave. To capture the tongue image, an ultrasound transducer should be placed beneath the chin during the image acquisition to acquire the most applicable view of the tongue contour. Ultrasound waves pass through the chin tissues in between the hyoid and mandible bones to reach the tongue. The impedance mismatch between the tongue tissue and the

air causes a strong reflection of the acoustic waves, which allows us to detect the tongue structure by detecting the reflected acoustic waves.

The tongue is positioned deeply in the oral cavity, making it challenging to fully view the contour during sound production. The hyoid and mandible bones absorb some acoustic waves, which may block the view of the tongue tip and root. Moreover, the shadowing of jawbones and instability of the head-transducer position would add other obstacles to the experiment.

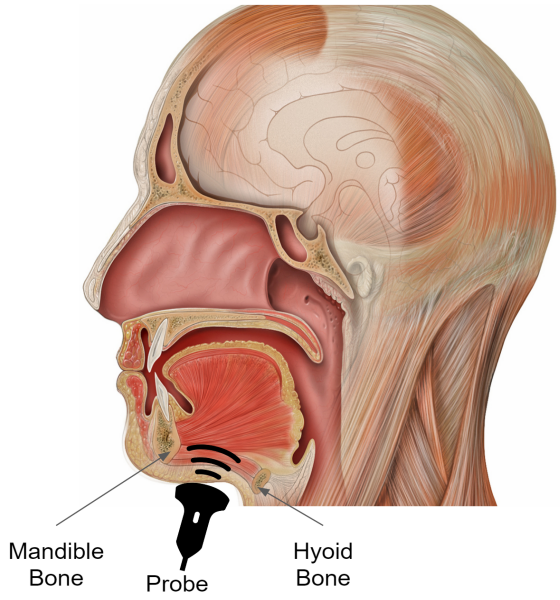


Figure 2.1: Overview of ultrasound probe placement beneath the chin. The head and oral cavity picture was modified from the original picture for the case, courtesy of Associate Professor Frank Gaillard, Radiopaedia.org, rID: 35836,[1].

Figure 2.2 shows the view of the tongue contour in the sagittal plane during the image acquisition using ultrasound. The final image of the tongue contour is presented on the ultrasound screen as a bright white concave arc. However, the the ultrasound system can detect the tongue image, but acoustic imaging is noisy by nature due to the low signal-to-noise ratio, and in the case of rapid tongue movements, there might be missing tongue parts in the image.

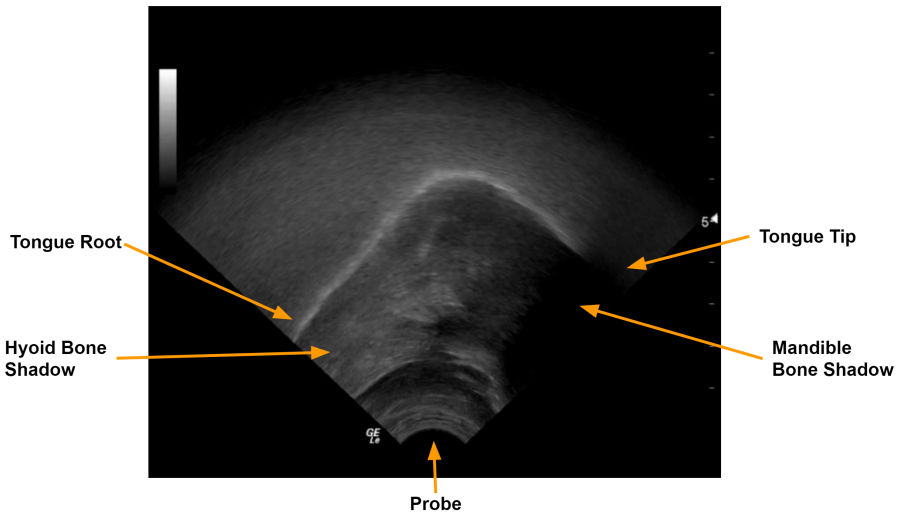


Figure 2.2: Ultrasound image of the tongue showing the tongue tip and root, Hyoid and Mandible shadows in the sagittal plane.

The typical ultrasound system configured with a microphone and the head-transducer support system arrangement [8]. Most of the image acquisition missing data are caused by ultrasound probe misalignment, losing the contact between the transducer and the skin, and the lack of acoustic gel that matches the impedance between the chin-transducer tip [96]. To alleviate image acquisition challenges, different measures must be taken into account. A skilled ultrasound specialist shall conduct the image recording session to properly acquire the image. During the session, it is recommended to use the head-transducer support system to stabilize the head and ultrasound transducer placement to maintain a fixed relative position between the transducer and the head.

A convex probe with a small and properly shaped tip area should be used to ensure the ultrasound waveform can pass through the bones and minimize the shadowing effect on the tongue tip and root. In addition, advanced signal and image processing techniques should be used to post-process and enhance the final image to ensure the data are clean and ready for analysis. In order to analyze and interpret speech further, the system records the sound of the speaker in parallel with the acquisition of the images.

2.3 Tongue contour tracking techniques in ultrasound

This section is a review of the tongue contour tracking methodologies in ultrasound images. There are two main subsections that categorize the tracking algorithms: first, traditional image analysis techniques for tongue contour tracking that review the non-training-based algorithms, which use a active contour model and a graph-based image analysis as core methodologies; second, machine learning-based techniques for tongue contour tracking to review the training-based algorithms that use machine and deep learning.

2.3.1 Traditional image analysis techniques for tongue contour tracking

Tongue tracking by ultrasound was addressed in early research by the cited works [97, 98]. However, the process was manual and required cautious user attention while handling the ultrasound transducer. To enhance the transducer guidance, metal pellets were used as a strong reflector to identify a few landmarks on the tongue surface. The landmarks were used as a reference to monitor tongue movement during swallowing by comparing the pellets placed on the tongue anterior and posterior segments to the Hyoid bone reference at different stages of movement. There are two main traditional methodologies used to segment the tongue: active contour model and shape consistency and graph-based tongue tracking models.

Active-contour-based methodologies

To automate tongue contour tracking, many researchers have relied on the active contour model [99, 100] as the base algorithm for most of the traditional techniques in tongue contour tracking. The active contour model is an active contour and energy-based method that adapts to get closer and closer to the object until reaching a certain threshold or energy constraints to fit the object boundary. The active contour model has been used widely in vision tasks such as the detection of lines, objects and subjective contours, and motion tracking. In the case of lingual ultrasound, the active contour model can be useful for interactively segmenting a tongue contour by applying certain user-imposed constraint forces to localize the tongue features of interest. Examples of the first attempts to use active contours

for tongue tracking tasks were provided by [101, 102] and [103], which were made by the same authors and improved consequently.

An adaptive active contour model was introduced by [101]. The authors collected 2D ultrasound images and used a head and transducer support system to stabilize the ultrasound transducer. In the first frame, a human expert selected a few candidates of the contour points to generate the initial tongue contour to initiate the active contour model. For the following frames, the researchers proposed an adaptive model that estimated an optimized contour that matched the tongue contour edges on each frame. Finally, the algorithm implemented a post-processing technique to enhance and refine the extracted contours.

The cited work in [102] followed the same process as the work in [101] and extended the work using different constraints to test it in speech and swallowing applications. The authors in [102] showed an improvement in the model performance by minimizing the computational cost to make it more flexible for a variety of different tasks.

Similarly, the algorithm proposed by [103] required an initial input from an expert to delineate the tongue contour on the first image frame to ease the active contour model optimization of the energy constraints that enforced the detection of tongue contour edges in the desired region of interest. Subsequent video frames were processed by adapting the initial contour edges to match the tongue deformation. External and internal energy functions were suggested to optimize the tongue contour's external edges and concavity, respectively. Although the methodology showed some success in tongue contour detection, its performance dropped drastically in the case of noisy images due to its sensitivity to speckle noise. Moreover, in the case of rapid tongue movements, the external energy function could fail to adapt the edges and match the tongue boundaries' deformation to the new position at the next frame. This, unfortunately, limited the ability of this methodology in real-time processing as it could fail suddenly during the video processing in real time.

Publicly available software EdgeTrack [13] proposed an improvement to the mentioned work in [103]. EdgeTrack implemented an enhanced methodology for the active contours that incorporated the gradient, local image information, and object orientation, unlike the classical methods that relied only on the gradient information [13]. This improvement optimized the contour's lower boundaries and rejected any undesirable edges unrelated to the tongue. EdgeTrack software

had a few technical limitations, and like any other deformable models, it could misidentify the true tongue contour's edges. EdgeTrack did not have any preprocessing capability, reducing the active contour model's efficiency as it is sensitive to noise. The software program could not process a long video sequence with more than 80 frames, limiting it to short recordings. This is not beneficial in the case of long speech processing sessions or a real-time analysis. EdgeTrack was computationally expensive because the algorithm relied on complex optimization techniques. In some cases, when there was a rapid movement during the speech, the tongue contour had a visible deformation that looked like a concave arc; the software tool failed because it did not use temporal smoothness in the minimized internal energy function. EdgeTrack results were validated by two experts who delineated the tongue contour manually. The mean sum of distances (MSD) accuracy measure was used to compare the results between EdgeTrack and manual ground truth data. The reported results were in the range of 1.83–3.59 mm for the MSD.

The multihypothesis approach [15] combined the traditional motion model, active contour model, and particle filter to track the tongue contour. The first step toward building the algorithm was by deriving a motion model based on manually prelabelled images. Next, tongue contours were extracted and then normalized with respect to the length and position. Following that, a principal component analysis (PCA) and mean shape were estimated, then the covariance matrix was computed by using the information from the tongue motion information such as the scale, shape, and position.

The active contour model used in [15] required to be initialized to process the tongue tracker by manually identifying points on the contour at the first frame to segment the tongue. After that, the particle filter was created by copying the segmented contour for a defined number of so-called particles. Next, a multihypothesis approach was created from each copied particle of the previous frame based on the derived motion model of the tongue scale, position, and coarse shape. The derived tongue contour model was then adapted using the active contour model to fit the tongue contour accurately. A band of energy-optimized constraints was used to choose the best particle by ensuring that the tongue contour was below the bright white arc on the tongue's upper surface. Two groups of subjects with Steinert's disease (a form of myotonic dystrophy that causes slow speech, distorted vowels, and consonants) and healthy subjects were used to validate the

research study. The reported accuracy was 1.69 ± 1.10 mm for the mean sum of distances (MSD). However, the approach claimed that it was not highly dependent on the training data. The segmentation accuracy was still dependent on the number of particles, which increased the active contour model's computational complexity [15].

To fully automate the tongue contour extraction without using training data or human interaction, some researchers designed multistage techniques [16] . Unlike other semiautomated methodologies such as those in [13, 14], and [104], which required human interaction in the first frame, this methodology initiated the active contour model by automatically deriving candidate points on the tongue contour. These points were identified by applying the phase symmetry method for image enhancement. Then, the image was skeletonized, and data points were clustered to select the best candidate points. These candidates were used as initialization points for the algorithm. The accuracy improved by implementing two methodologies for algorithm resetting or reinitialization in a frequent and timely manner order. According to the results, the measured mean sum of distances (MSD) accuracy measure was similar to that of other semiautomated techniques. They claimed that the MSD was 1.01 mm and 0.63 mm for their fully automated and reinitialized techniques, respectively. The reported results were highly accurate with some frames, but this may not be easy to achieve when processing videos in real time.

However, relying on the active contour model for tongue tracking in ultrasound images is error-prone and maybe not the most efficient technique. In some cases, it can lead to ultimate failure due to the number of constraints needed for the model adaption, which is difficult to predict for all cases accurately. Although the approach in [16] proposed a novel methodology for automating the process of identifying the active contour initialization and reinitialization parameters, this was still not enough to produce highly accurate results in a global and generalized context. There are many variations in ultrasound imaging modalities that produce different imaging qualities, making it difficult to track the tongue contour using the same active contour model constraints.

The similarity-constrained active-contour-based methodology for tongue tracking proposed in [105] suggested a technique that coped with the tongue contour tracking errors and missing data based on the tongue shape from previous contours to minimize the effect of missing data. In order to deal with the accumulated error during the continuous tracking of the tongue contour over a video sequence,

a complex-wavelet image similarity index (CW-SSIM) was proposed to reinitialize the tongue tracker automatically. This algorithm showed an advancement compared to traditional techniques by handling missing data and using an automatic re-initialization. However, it was still based on the active contour, which is error-prone and sensitive to noise. Too many constraints would enhance the model accuracy but increase the computation cost. The best-reported results using similarity constraint + CW-SSIM were an MSD of 0.9912 ± 0.2537 mm.

As mentioned before, all methodologies that are based on the active contour may suddenly fail and the tongue tracker would stop. An initializer, either manual or automatic, is needed to enhance the accuracy of tongue tracking. The researchers in [106] conducted a comparative study on the effect of an automatic re-initialization technique to enhance the well-known traditional image segmentation. The automatic re-initialization enhanced the results from an MSD of 5–6 pixels to about 4 pixels (1 pixel = 0.295 mm). The MSD accuracy results without the need for automatic re-initialization for the well-known tongue tracking tools EdgeTrack and TongueTrack were 7.06 ± 2.77 pixels and 5.59 ± 3.04 pixels, respectively. The MSD accuracy after using the automatic re-initialization was 3.46 ± 1.04 pixels and 3.60 ± 0.96 pixels for EdgeTrack and TongueTrack, respectively.

Shape consistency and graph-based tongue tracking methodology

Researchers derived an active appearance model to predict the tongue contour shape on ultrasound images in [107]. The active appearance model was inspired and estimated using a manual delineation and extraction of the tongue contour from tongue X-ray images. The results were compared to those of EdgeTrack [13] and the constrained active contour model [108], which combined ultrasound, EMA, and recorded voice to predict the tongue shape. The work in [107] showed an improvement in root mean square error compared to that of [13, 108]. The active shape model (ASM) was also evaluated and used in [98]; the authors showed that the ASM was efficient and powerful for phonological applications. It was able to capture the tongue motion variation by capturing the temporal information. It was also useful for either automated or semi-automated techniques.

Lingual ultrasound tracking was introduced in another well-known software called [14] TongueTrack, which could process a sequence of 500 frames. The method-

ology considered contextual information and advanced optimization techniques to estimate unpredictable tongue motion. The reported accuracy was 3 mm, making it acceptable for segmentation purposes. The tool used a higher-order Markov random field energy minimization framework. The results were validated with the ground truth data from two different groups of 63 acoustic videos [14].

The process of TongueTrack required an initial human interaction by manually delineating a few points on the first tongue contour to be used to initialize the algorithm. After that, the delineated points were fitted by using a curve-fitting polynomial function to build a continuous and smooth contour. Next, a solution-space label set was created by generating an estimation model for the dynamic tongue motion. This label set was used to compare each contour with the minimized Markov random field energy module in each subsequent frame. It processed it iteratively until reaching a predefined threshold; it was predefined as 2 mm in [14]. The tool obtained good results, but it had a few drawbacks. The software tool could not process long video frames. At the same time, the algorithm optimizer might not converge properly, leading to a sudden failure in tracking progress as it required 20 iterations to optimize nine parameters. Moreover, the algorithm needed a manual re-initialization by delineating the tongue contour by hand, limiting its efficiency for real-time processing.

Tongue contours are also tracked in ultrasound images by using graph-based analysis of the temporal and spatial information during speech [109]. Spatial information is essential to extract tongue features from each image on a single frame. At the same time, the temporal resolution is necessary to predict the interrelationship between the entire sequence of image frames extracted from the video session of the speech. The tongue tracker was implemented as an optimization problem using a Markov random field energy minimization. The algorithm enforced temporal and spatial regularization constraints to ensure tongue tracking reliability.

In the landmark-based tongue contour tracking [104], the tongue shape was predicted based on the position of a few pellet plates used as landmarks on the tongue surface. The landmarks were extracted from the available articulatory database. The available landmark positions were smoothed using the spline function and compared to the ground truth data extracted by ultrasound images. Tongue contours extracted by ultrasound helped to identify the optimum number of required landmarks to get the desired accuracy of 0.2–0.3 mm for any future use.

Another research study coped with the tongue tracking problem by modelling

it as a biomechanical method [110]. The methodology was initialized by manually drawing a closed contour around the external and internal edges of the tongue. The Harris feature detector was used to identify the one hundred most significant corners or edge features. The detected points were sorted in descending order based on the quality of the feature. An optical flow algorithm was then used to estimate each point's displacement in the consequent frames. The corner feature displacement estimation was approximated only in the neighbour pixels (around 15–20 pixels) to minimize the displacement error in case of any missing data. In order to minimize the uncertainty of the estimated features, a covariance matrix was computed. The accuracy was measured by the mean sum accuracy, which was reported between 0.62 mm and 0.97 mm. However, the study faced many challenges. The algorithm required many parameters and constraints to be computed in order to estimate the displacement. Relying on the Harris feature detector may not have been efficient, especially in the case of rapid tongue movement, missing details, or extreme deformation, as it was almost impossible to guarantee that the same detected corner features were visible in the next frame within the neighbourhood pixel constrains.

An interactive approach for lingual ultrasound segmentation that incorporated four stages from pre-processing to the segmentation and post-processing analysis was introduced in [8]. In the first stage, and unlike other methodologies that ignored an essential part of image denoising, the thesis implemented novel denoising techniques by using a combined curvelet transform and shock filter. In the second stage, the thesis derived an interactive model that predicted the tongue area of interest to minimize the computation complexity and contour tracking error. The third stage focused on tongue contour extraction and smoothness. The fourth stage proposed a new technique that transformed the extracted tongue contour from an image state to a continuous signal which resembled a full video for all frames. The advantage of this technique was that it enabled the researcher to extract a unique signature of each sound; this could be beneficial for training a machine learning model on sound pattern recognition. The tongue contour segmentation results were validated and compared to ground truth data. The mean sum of distances (MSD) was 0.955 mm.

2.3.2 ML-based techniques for tongue contour tracking

One of the early attempts to use deep learning for automatic tongue extraction was made by [111]. Their methodology, Autotrace, was implemented using a translational deep belief neural network (tDBN), which was based on restricted Boltzmann machines (RBMs). The network was trained based on human-labelled and generated sensor data. The hybrid data training methodology was efficient for improving tongue contour segmentation accuracy. However, there were discrepancies in the segmentation of some image frames and model-segmented tongue-unrelated parts. The results were validated by using a five-fold cross-validation, and the reported accuracy was measured by an average mean sum of distances (MSD) of 2.5443 ± 0.056 pixels (1 pixel = 0.295 mm [13]). The algorithm segmentation capabilities were fair enough; however, a post-processing algorithm was needed to refine and enhance the final tongue contour segmentation.

To improve Autotrace [111], researchers in [112] proposed a new technique that automatically labelled the tongue contour, followed by training the algorithm in two phases. Using a deep autoencoder, the algorithm learned the relationship between the extracted contour and the original ultrasound image. By using the training data, the algorithm was able to reconstruct the tongue contour from ultrasound images without human intervention. The results were validated by comparing the average mean sum of distances between the hand-labelled and the deep-learning-extracted contours. The average MSD was reported as 1.0 mm, making it applicable to lingual ultrasound applications.

Based on the principal component analysis (PCA) and a neural network, an automatic algorithm was designed to segment the tongue contour [113]. The PCA-based feature extractor, Eigen Tongue, was used to extract the tongue contour features from the ultrasound images. The visual features of the extracted Eigen Tongue were processed using an artificial neural network based on the PCA feature model. The model was evaluated by using 80 annotated images from nine speakers. The average error measured by the MSD was reported to be around 1.3 mm.

Typical convolutional neural networks were used to classify the tongue gesture from B-mode ultrasound images on the midsagittal plane in [114]. The researchers used data augmentation to increase the size and versatility of the data, which increased the algorithm's performance. The reported accuracy results for the classi-

fication task were 76.1%. Further improvements were suggested as future work. The recommended improvements were in the model optimization or combining the methodology with a hybrid technique such as the ensemble method.

The well-known U-Net architecture [115] was used by [116] to automatically extract the tongue contour in ultrasound images. The algorithm was trained by using 8881 human-labelled images collected from three subjects. The results were validated by using the Dice score, which was 0.71. Relying on the Dice score only for validation is not enough. More validation is needed for their methodology, such as the mean sum of distances (MSD) measure, which has become a de-facto standard in the lingual ultrasound accuracy measures. The MSD provides a reliable measure that considers the variation of the tongue contour length, which normalizes the sum of distances over the tongue contour length. To further enhance the performance, it might be needed to use a hybrid technique and larger dataset.

To automate tongue segmentation, a convolutional-neural-network-based architecture was utilized in [117]. They compared the efficiency of using the U-Net [115] and Dense U-Net [118] architectures to extract the tongue contour. These architectures have become de-facto models of biomedical image segmentation and gained a wide popularity in the field. The results showed that Dense U-Net was can be generalized for a wide variety of datasets. At the same time, the standard U-Net architecture could perform the tongue extraction task faster. After extracting the tongue contour, it had to be post-processed. In the post-processing stage, the output was fed into a probability heat-map model, where the intensity of each pixel corresponded to the probability of each part of the tongue [117]. A 50% threshold was applied to filter out any undesired predictions. The remaining output was utilized to reduce the segment thickness. Following that, the results were smoothed and interpolated using the UnivariateSpline function in the SciPy package in Python. The final output was a hundred points to represent the predicted tongue. The algorithms were evaluated using the MSD for the 17,580-frame dataset. The reported MSD results for the 32×32 data size were 5.81 mm and 5.6 mm for U-Net and Dense U-Net, respectively. The research also showed that data augmentation and the loss function significantly affected model performance other than stacking more layers.

Two deep learning architectures were designed, BowNet and wBowNet, to extract the tongue contour from ultrasound in [119]. With the integrated multi-scale contextual information, the decoding–encoding model had the ability for global

prediction. The dilated convolution had the local searching capability of preserving image features more than standard convolution, making it valuable for medical imaging applications to retain fine image details. The two architectures enhanced the final prediction results by combining the local and global searching. The mean sum of distances for BowNet and wBowNet compared to the grey-scale ground truth images was in a range of 0.2874–0.4014 in pixels for BowNet and 0.1803–0.3588 pixels for wBowNet. However, the reported results appeared to be almost perfect, which is not easy to achieve in the case of a complex analysis of lingual ultrasound. The researchers need to provide more information about the data validation in a generalized clinical context by using a dataset from a different source.

A simple approach to extracting the tongue contour by training a deep network on landmarks annotated on the tongue contour was developed in [120]. These landmarks were automatically and randomly selected on different points by using annotation software. The model architecture was called TongueNet, and the results were validated by the mean sum of distances which achieved 4.87 pixels.

Using U-Net and the lighter version of sU-Net in a thesis work, a deep learning approach was implemented to segment tongue contours [121]. In their thesis, the researcher emphasized the validity and performance of deep learning models to segment the tongue contours from ultrasound images. However, they suggested that the deep learning model they used only focus on the spatial information on a single image frame without considering the temporal information that handled the full speech in the video sequence. The thesis [121] also discussed the limitations of their deep learning model in their generalization capability of feature extraction, as they inherited the non-generalization of convolutional neural networks (CNN) models, which is the core of a deep learning model such as the U-Net architecture. The thesis suggested using data augmentation to enhance the model training by considering the variation and image transformation to handle different cases at different scales.

A denoising convolution autoencoder (DCAN) model to process B-mode ultrasound images was investigated in [122]. The model reported being able to extract image features due to its ability to denoise and retain the resolution of the reconstructed input from the ultrasound. It was tested on reconstructing ultrasound images in speech-related applications. The research compared the DCAN to other three well-known autoencoder architectures, the deep autoencoder (AE), the denoising autoencoder (DAE), and the convolutional autoencoder (CAE). The re-

ported result showed that the DCAN had a 6.17% error rate in identifying words in a silent-speech recording test [122].

Researchers implemented a novel technique that harnessed the spatial–temporal analysis to predict future tongue movement based on a short recording of the past tongue motion in [123]. The research used a combination between a convolutional neural network (CNN) and long short-term memory (LSTM), which was called ConvLSTM. The advantage of this combination was that the CNN had the ability to segment tongue contour in each image frame to extract spatial information. However, it could not process the temporal information of ultrasound image sequence frames. On the other hand, LSTM was used in processing data sequence in one dimension, making it efficient for temporal information data prediction, but at the same time, it was unable to handle images in two dimensions (2D). The ConvLSTM could handle image data in 2D and predict future data based on the history of tongue motion. The ConvLSTM results outperformed the three-dimensional convolutional neural network (3DCNN) in predicting future tongue contours. The ConvLSTM was able to predict the future nine frames based on data from the previous eight frames. We believe this algorithm was not only important for data prediction of tongue contours, but it might be helpful for generating more data that are close to real data to train larger deep learning algorithms such as a Transformer model or a graph neural network.

An algorithm combining an image-based segmentation model, U-Net, and a shape consistency regularizer was proposed by [124]. The combination provided a solution to the missing data in ultrasound images by predicting the information based on the consideration of the sequential information of the shape regularizer. The regularizer was derived based on the similarity between adjacent image frames. The results were validated by computing the MSD of the tongue contour data segmented by the U-Net algorithm using different loss functions. The quantitative validation showed that the combination between the regularizer and cross-entropy loss (CE) obtained the best results among the other compared losses such as the Dice coefficient (DC) or the active contour loss (AC). The CE+regularizer reported having an MSD of 2.243 ± 0.026 mm.

To improve the well-known U-Net architecture, researchers proposed a tongue contour segmentation algorithm called wUnet [125]. The main modification of wUnet was replacing the skip connection in typical U-Net with a VGG19 block. The researchers claimed that the new algorithm surpasses U-Net by passing more

information to the decoder to compensate for the information loss during the convolution within the encoder. The wUNet validation results showed an MSD of 1.18 mm compared to 2.26 mm in the U-Net architecture.

A system based on a deep learning technique was designed to predict silent speech using ultrasound images in [126]. The system was trained on audio features recorded synchronously with ultrasound images using a deep convolutional neural network. The system was designed to predict the speech sound from the silent speech based on the training data. This methodology could be beneficial for human–machine interaction in smart devices.

To update an older silent-speech benchmark study [84], the work [83] used a deep learning approach for the same benchmark. The new study used a deep autoencoder to train the collected dataset from acoustic tongue and lips movement videos, which were collected at the same time.

The research [19] used ultrasound videos to extract tongue features using deep learning. The dataset was collected from 82 speakers and trained using the Kaldi speech recognition toolkit [127]. In terms of speech analysis, the research suggested two methodologies. The first one was the utterance or speech duration, which was measured based on the syllable rate. The second one was the articulatory area, which was measured by estimating the convex hull area, which was the area under the tongue contour spline that formed a convex-like shape when extracted from the ultrasound images using the MTracker tool [116]. Following that, a postprocessing was performed by the isolation forest method [128]. The research found that the silent articulation exhibited a longer time compared to the model speech.

Chapter 3

Vision-Transformer Architecture Design and Applications in Digital Health

This chapter includes the problem formulation, Vision-Transformer architecture design and study of the feasibility of ViT in medical ultrasound tongue tracking by conducting a survey of different applications in digital health. Problem formulation, feasibility study and architecture design is the most important step in any machine learning project. It is essential to decide whether the problem can be handled with AI or not and what is the appropriate model for it.

3.1 Background

The COVID-19 pandemic provided an example of how AI can help scale the system in emergencies when the medical staff is limited or has safety concerns. AI algorithms are used widely in digital medicine solutions, mainly in image and text recognition tasks to analyze medical data stored in clinical information systems to generate medical reports and assist other technical operations like robotic surgery. Among the various AI-assisted tools for analyzing medical images, Vision Transformer (ViT) emerges as the state-of-the-art algorithm that replaces or combines traditional techniques like CNN. This article aims to discuss the foundation applications of ViT in digital health.

Vision-Transformer (ViT) [2, 129] is a type of neural network used in computer

vision tasks to process images [130]. ViT backbone is based on the self-attention mechanism and typical Transformer used in natural language processing. ViT is introduced to address the limitations in image processing of other common machine learning architectures like convolutional neural network (CNN) [131], Recurrent Neural Network (RNN) [132] and even the traditional Transformer that is proposed for language model [2, 133]. ViT provides a strong representation of image features and trains the data using fewer computational resources compared to the convolutional neural network [2].

CNN is widely adopted in the machine learning field, and it is well-suited for feature extraction on specific local regions, but it lacks the ability to capture the contextual relationship between image features in the global context. On the other hand, ViT applies an attention mechanism to understand the global relationship among features.

RNN is used to get inferences about the sequence-to-sequence relationship and memorize some data from the past. It requires memory, but it is not ideal for extracting image features compared to ViT or CNN. The Bidirectional Encoder Representations from Transformers (BERT) was developed by Google to process language models [134] based on the attention mechanisms [135]. BERT is efficient for processing sequence-to-sequence models and requires larger memory compared to RNN or Long-Short-Time-Memory (LSTM) [136].

BERT has a limitation in processing data in image format. BERT works only for flattened data in a sequence shape. To solve the BERT image analysis issue, ViT splits images into patches and flattens them to analyze them as a linear sequence [2] in a parallel processing mechanism.

The application of Vision-Transformer in medical imaging can be classified into main categories. Segmentation, classification, reconstruction, prognosis prediction and application in telehealth like report generation and security.

3.2 Vision-Transformer Architecture

This section discusses the core principle and foundations of the Vision-Transformer based on the attention mechanism. The Vision-Transformer architecture is broken into a hierarchy of different functional blocks that will be explained in the following sub-sections. Fig. 3.1 depicts the typical Transformer architecture that is proposed by [135], based on the attention mechanism. The detailed explanation of

the architecture are discussed in the following subsections.

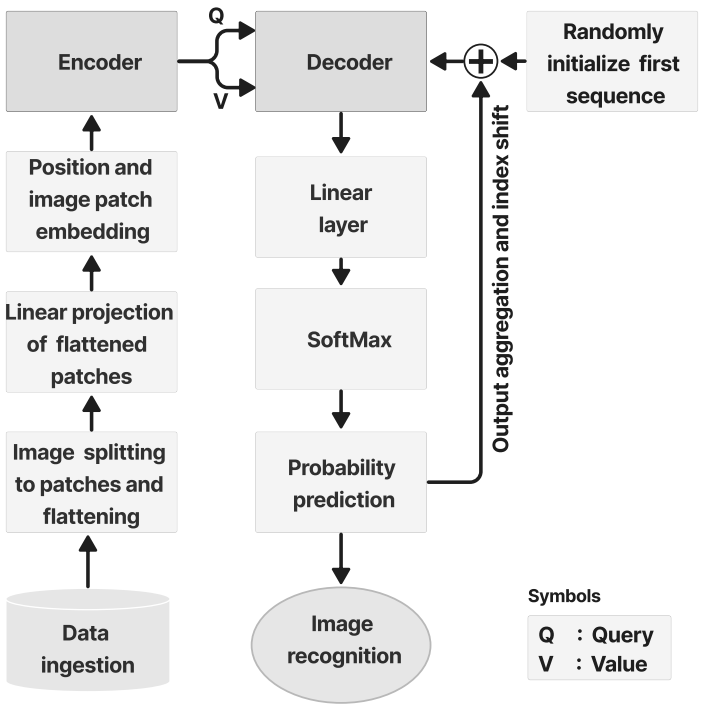


Figure 3.1: Illustration of a high-level block diagram of the Transformer architecture. Q is query, and V is value attributes in transformer model [2].

Researchers proposed different modifications to make the typical Transformer [135] design applicable for applications other than Natural Language Processing (NLP) tasks. The changes focused on the design framework of encoder-decoder blocks in the Transformer architecture. The Transformer is useful in the vision tasks by splitting the image into patches and flattening them into sequence forms to be processed like times series data which is more suitable to the Transformer nature. To ensure that we can reconstruct the image without any loss in the data, positional encoding is utilized for the embedded features in a vector shape. The embedded features are fed into the encoder for the image classification task and then classified using multi-layer perception [2]. On the other hand, for the segmentation task, the Transformer is combined with the convolutional neural network either in the encoder stage like the TransUNet architecture [137] or in both

the encoder-decoder stages like the Ds-TransUnet [138].

3.2.1 Encoder Architecture

The typical encoder is composed of a stack of (N) identical layers. Each layer contains two sub-layers or two stages. The first sub-layer performs the Multi-Head Self Attention (MSA). The second sub-layer normalizes the first sub-layer output and feeds it into the Multi-Layer Perceptron (MLP), which is a type of feed-forward network. Fig. 3.2 depicts the typical encoder architecture [135].

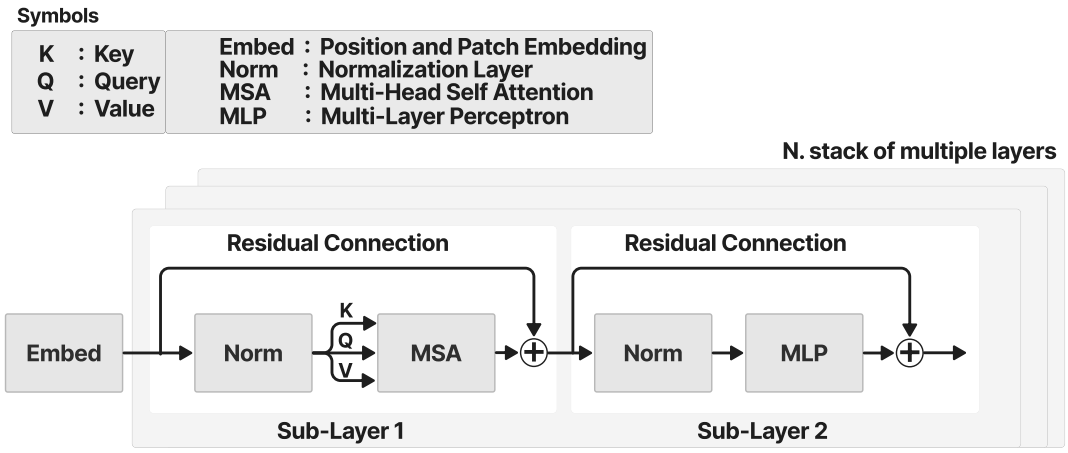


Figure 3.2: Encoder block in the Transformer architecture [2].

3.2.2 Image Patches Embedding

Due to the computer memory limitation, it is difficult to process the whole image simultaneously. For this reason, the image is divided into different patches and processed sequentially. To give a meaningful analysis of each image patch, each patch is embedded into a set of feature values in the form of a vector.

The idea of image patch embedding in ViT is inspired by the word embedding in work proposed in [139]. The feature vectors are then graphically visualized in an embedding space. Presenting the features in the embedding space is beneficial to identify the image patches with closer similarity features [140]. We can simply measure the distances between each feature in the features map and identify the

degree of similarity [141].

Fig. 3.3 depicts the feature embedding process that starts by creating an embedding layer from the embedding vectors of each input feature. Random embedding values are initially assigned and updated inside the embedding layer during training. In training, the features with more similarity are getting closer and closer to each other in the embedding or latent space. This is very important to classify or extract features that are similar to each other. Without knowing the position of each feature, it is difficult to find the relationship between them. In medical imaging applications, positional encoding and feature embedding are beneficial to give an accurate feature selection for the desired use case.

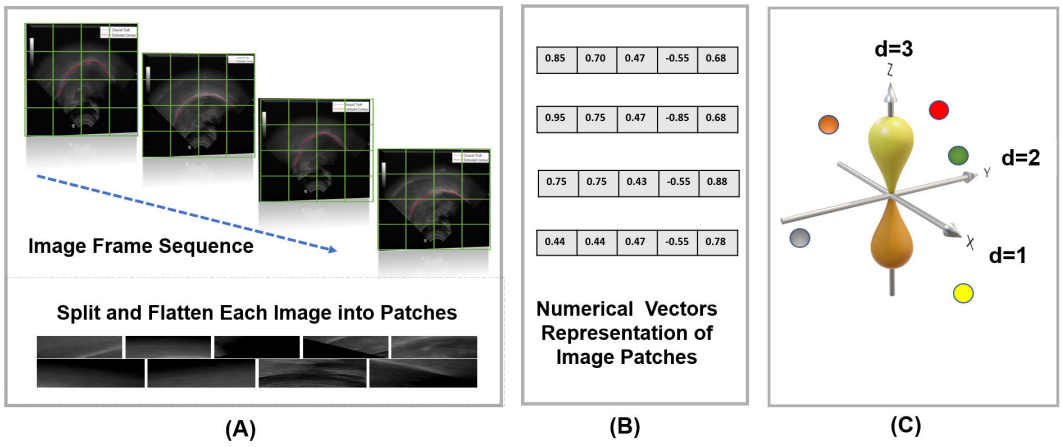


Figure 3.3: (A) Illustration of splitting ultrasound images into patches and flattening them in a linear sequence. (B) Image patch vectorization and linear projection. (C) Patch embedding is represented in multi-dimensional space.

3.2.3 Positional Encoding

The Transformer model has the advantage of simultaneously processing inputs in parallel, unlike the well-known Long-Short-Term-Memory (LSTM) algorithm [142, 143]. However, parallel processing is challenging because there is a chance of losing some information due to the inability to reconstruct the processed sequences into their original position. Positional encoding is proposed to solve this issue and encode each feature vector to its accurate position [135, 144]. Fig. 3.4 depicts the positional encoding process for the feature representation.

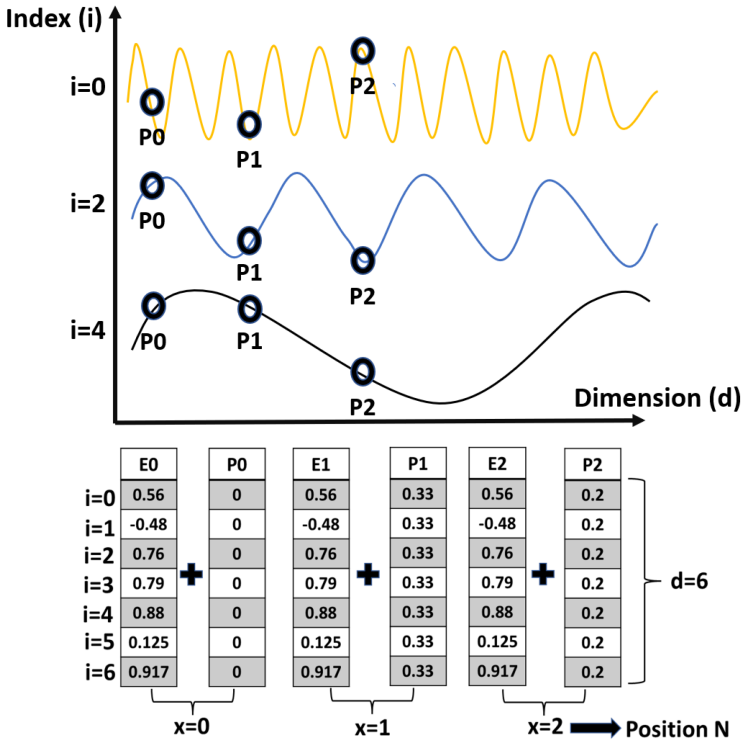


Figure 3.4: Positional encoding for the feature representations. Top: Illustration of sinusoidal representation for the positional encoding (P0-P3) at different indices and dimensions. Bottom: Illustration of vector representation for the positional encoding and feature embedding; P is the position encoding and E is the embedding vector.

The feature vector and positional encoding values are added to each other to form a new vector that is represented in the embedding space. In this paper, we use the *sine* and *cosine* functions as examples to derive the positional encoding values at different frequencies. Eq.(3.1) and Eq.(3.2) are the *sine* and *cosine* functions [135].

$$P(x, 2i) = \sin(\frac{x}{10000^{2i/d}}) \tag{3.1}$$

$$P(x, 2i+1) = \cos(\frac{x}{10000^{2i/d}}) \tag{3.2}$$

where P is the positional encoding, d is the vector dimension and x is the position, i is the index dimension. The sinusoidal function is beneficial for encoding

the feature position in the embedding space by using different frequencies ranging from 2π to 10000. In Eq.(3.1) and Eq.(3.2), the frequencies resembled as the index dimension (i) [135].

3.2.4 Multi-Head Self Attention (MSA)

Fig. 3.5 illustrates the multi-head self-attention (MSA) process. MSA calculates the weighted average of feature representations based on similarity scores between pairs of representations. The input sequence X of L tokens or entries with dimension d, where $X \in R^{L \times d}$ is projected using three matrices $W_K \in R^{d \times d_k}$, $W_Q \in R^{d \times d_q}$, $W_V \in R^{d \times d_V}$ with the same dimensions to derive the representation of the features. Eq.(3.3) shows the formula to derive Key, Query and Value.

$$K = XW_K, \qquad Q = XW_Q, \qquad V = XW_V \tag{3.3}$$

The final Embedding Layer that includes Position Encoding is copied into the three linear layers Key (K), Query (Q) and Value (V). To derive the similarity between the input features, a matrix multiplication between the Key and Query is performed using Self-Attention. The output is then scaled and normalized using *SoftMax*. The process of Self-Attention [130], is explained in a few steps as follows:

1. Calculate the score from the input of Query and Key.

$$S\text{=}\ Q\ K^T \tag{3.4}$$

2. Normalize the score to stabilize the training.

$$N_S\text{=}\ S\sqrt{d} \tag{3.5}$$

3. Calculate the probabilities of the normalized score using SoftMax.

$$P = SoftMax(N_S) \tag{3.6}$$

4. Compute the self-attention filter by multiplying P and V.

$$Self Attention = PV \tag{3.7}$$

The multiplication output of K and Q is scaled by the square root of the input vector dimension and then normalized by the SoftMax function to produce the probabilities. Eq.(3.8) shows the formula for the Softmax function, where x is the input data point. While Eq.(3.9) derive the attention filter.

$$SoftMax(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{3.8}$$

$$Self\ Attention(Q, K, V) = SoftMax(\frac{Q.K^T}{\sqrt{d}})V \tag{3.9}$$

The output probabilities from SoftMax and the value layer are multiplied to produce the desired output with more attention or emphasis on the desired features and filter out unnecessary data. The intuition behind the Multi-Head is concatenating the results from different attention filters, each one focusing on desired features. The self-attention process is repeated multiple times to form Multi-Head Self Attention (MSA). The final output of the concatenated multi-head self-attention is passed through a linear layer then it is resized to the single-head size. Eq.(3.10) shows the MSA formula.

$$MSA(Q, K, V) = C(h_1, \cdots, h_n)W_0 \tag{3.10}$$

Where C is the concatenation of multi-heads. W_0 is the projection weight. Q, K and V are the Query, Key and Value, respectively. While h resembles each head in the Self-Attention process and is replicated to n times. The number of replication depends on the number of attention or the desired features that are required to extract the required information. Fig. 3.5 depicts the process of Multi-head Self Attention in the Vision-Transformer architecture. Detailed information about the scaled dot product between K, Q , V is also depicted.

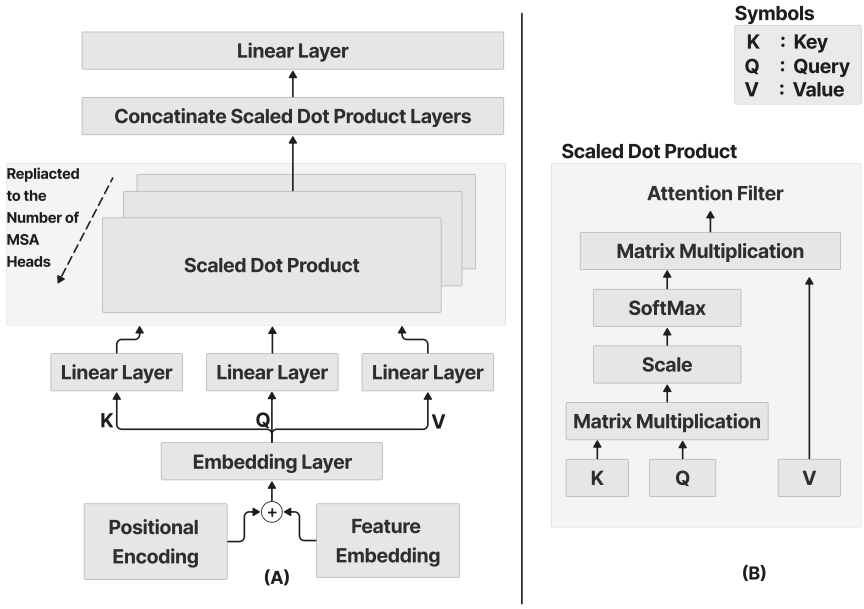


Figure 3.5: Overall MSA process. (A) illustration of MSA process with several attention layers in parallel. (B) Scaled dot product.

3.2.5 Layer normalization and residual connections

Residual connection is essential to feed the output from the position encoding layer into the normalization layer directly by bypassing the multi-head self-attention layer [145]. The residual connection is essential for knowledge preservation and to avoid vanishing gradient problems [146, 147]. The multi-head attention layer is vital for extracting useful features from the input but has disadvantages. It may disregard helpful information that is less weighted in the attention filter. Minimizing the value of feature weight may cause a vanishing gradient during the model training stage. Vanishing gradient happens when the gradient of the loss function is depleted to be equal or almost zero while optimizing the weight in the back-propagation algorithm. The residual connection feeds the information from early layers directly into the layers at the end of the neural network to preserve features and not forget important information.

The add and normalization layer [148] combines the input from position encoding and the multi-head self attention layer and then normalizes them. The normalization layer is essential during the training to make the convergence of the

loss faster and more stable. The normalization can be derived by standardizing the activation of the neurons along the axis of features. Eq.(3.11) and Eq.(3.12) are the statistical components of the layer normalization over all the hidden units in the same layer [148].

$$\mu^l = \frac{1}{H}(\sum_{i=1}^{H} a_i^l) \tag{3.11}$$

$$\sigma^l = \sqrt{\frac{1}{H}(\sum_{i=1}^{H} a_i^l - \mu^l)^2} \tag{3.12}$$

where a_i^l is the layer normalized value of the summed input features along i^{th} hidden units in the l^{th} layers. Similarly, H is the total number of hidden units in the layer. While μ is the mean or average of values of features along the axis in the normalization layer, σ is the standard deviation of the values of the features along the axis, and ϵ is the small number added to avoid division on zero.

3.2.6 Multi-layer perceptron (MLP)

Fig. 3.6 depicts the multi-layer perceptron (MLP) diagram, which is part of the ViT architecture. MLP is a type of feed-forward artificial neural network that combines a series of fully connected layers that includes input, one or more hidden layers in the middle, and output [149]. Fully connected layers are the type of layers in which each neuron's output is connected to all neurons in the next hidden layer. The diagram shows that Each neuron from the layer in the feed-forward neural network is connected to all neurons in the next layer through an activation function. Residual connection preserves knowledge from initial layers and minimizes the vanishing gradient problem. Typical MLP layers have input, output and hidden layers.

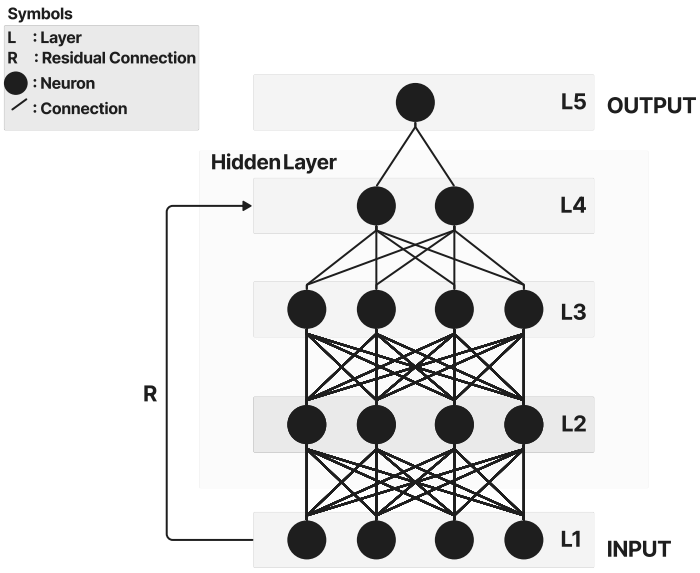


Figure 3.6: Multi-layer perceptron (MLP).

3.2.7 Decoder and mask multi-head self attention

Fig. 3.7 illustrates the decoder and mask multi-head self-attention in the ViT architecture to extract the final image. The decoder is stacked for N multiple layers as the same encoder number of layers. The decoder includes the same sub-layers in the encoder and the mask multi-head self-attention that is stacked to them. mask multi-head self-attention works similarly to multi-head self-attention (MSA), but it focuses the attention on the desired feature in position i to ignore any undesired feature from the embedding layer by mask-only features that are before i. This is important to get an inference from the relationship between different features in the embedding space and get a prediction from features that are relevant to the desired position.

The decoder gets V, Q and V as inputs. The value V is fed from the previous embedding space. The Query Q and Key K are fed from the encoder output. Inside the decoder, there are another MSA and normalization layers. This is the typical design of ViT. However, there are different modifications of the decoder-encoder design, but the core principle is the same; different architectures for different applications are discussed in Section 3.3.

The decoder output is flattened as a linear or dense layer in the image recognition task. Following that, $SoftMax$ is used to derive the probability of the weight of each neuron in the dense layer. The final probability is used to classify or segment the features based on the training data to detect the final object or image.

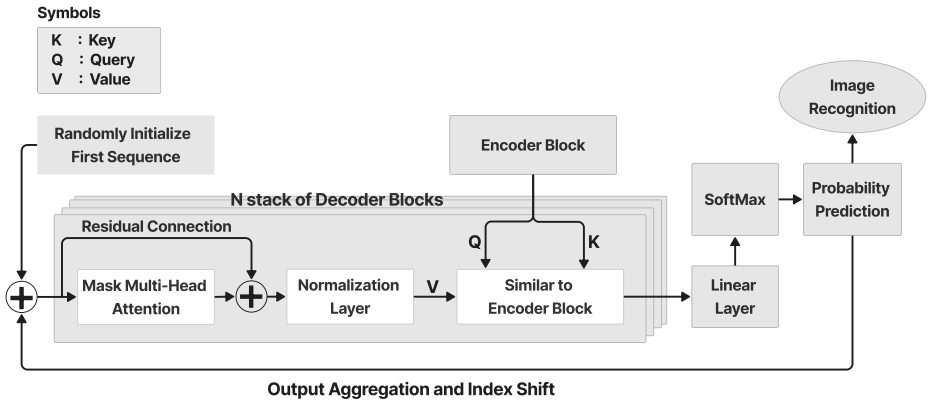


Figure 3.7: Decoder and mask multi-head attention block to produce the final image.

3.3 Application of the ViT in Digital Health

Computer vision and machine learning algorithms have been employed in different medical studies. It has bee used for brain and breast tumor [150, 151], histopathology [152], speech recognition [10, 8], rheumatologists [153], automatic captioning [154], endoscopy [155], fundus imaging [156] and telemedicine [157]. Vision-Transformer is emerging as the state-of-the-art in AI-based algorithms that use computer vision and machine learning for digital health solutions.

Fig. 3.8 depicts the distribution of the Vision-Transformer applications in the medical field according to the survey [3]. The primary use of ViT-based architectures is in the application of medical segmentation, detection, classification, report generation, registration, and other applications like prognosis prediction and telehealth.

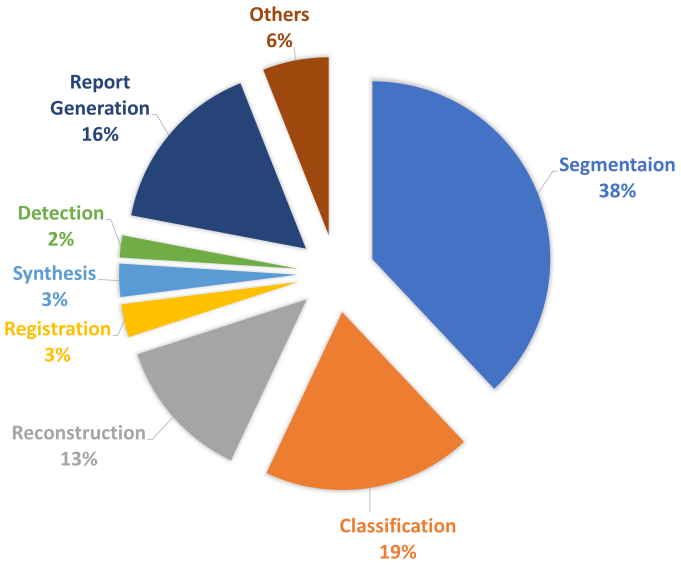


Figure 3.8: Distribution of medical image application of the ViT; results according to the survey [3].

3.3.1 Application of ViT in Medical Image Segmentation

One of the early attempts to use ViT in medical imaging segmentation is the TransUNet [137]. It is proposed to combine the capabilities of Vision-Transformer (ViT) and UNet [115] architectures. Unet is well known in the area of biomedical image segmentation, and it is efficient for object segmentation tasks and can preserve the quality of the fine image details after reconstruction. Unet inherits the localization ability of feature extraction in the convolutional neural network (CNN). While localization is essential in the segmentation task, it has limitations in the case of processing sequence-to-sequence image frames or even extracting global features within the same image outside a specific region. On the other hand, ViT has the advantage of processing sequence-to-sequence features and extracting global relationships between them. However, ViT has limitations in feature localization compared to CNN's abilities. TransUNet proposes a robust architecture to combine ViT and Unet in one model to take advantage of each technique's capabilities.

TransUNet is powerful in multi-organ segmentation in medical applications. Segmenting different objects is essential to analyze complex structures from mag-

netic resonance imaging (MRI) or computed tomography (CT) images. Fig. 3.9 depicts an image segmentation example of the abdomen from a CT scan using TransUNet and compares it with other techniques.

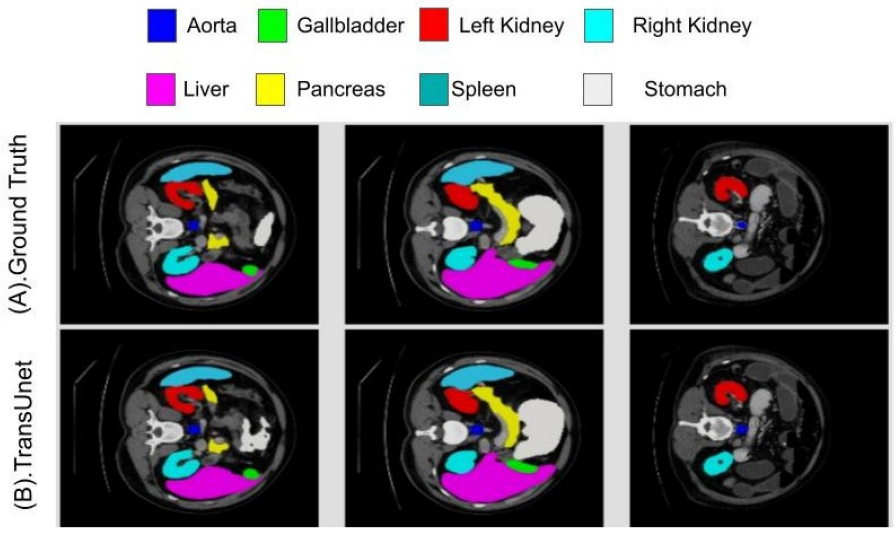


Figure 3.9: Comparison of TransUNet and Ground Truth using output segmentation results of different organs. (A) Ground Truth (Expert Reference) (B) TransUNet.

To further improve the TransUNet architecture, a Dual-TransUNet is implemented in work [138]. The main difference is that the Dual-TransUNet uses the Transformer in the encoder to extract features and the decoder to reconstruct the desired image. While the TransUNet only uses the Transformer in the encoder stage. Swin-Transformer [158] is another successful architecture to implement the vision-Transformer in combination with UNet [115, 159] in the medical imaging context.

The Vision-Transformer is also used in iSegFormer [160], which is proposed to make an interactive image segmentation of the 3D MRI knee images. 3D UX-net [161] can segment brain tissue from the whole body of the MRI scan. UNesT [162] developed a hierarchical Transformer using local spatial representation for medical image segmentation in different applications, including brain, kidney and abdominal multi-organ segmentation. In the same manner, NestedFormer [163] is proposed to segment brain tumours from MRI images. RECIST [164] uses the Transformer to automatically segment brain tumours to measure the size of the

lesion from CT images. On the other hand, GT U-Net [165] is used for tooth therapy by segmenting the tooth root canal from X-ray images. Colorectal cancer images can be segmented using FCN-Transformer [166] during Colonoscopy. ViT is also used in TraSeTR [167] in robotic surgery to assist the robot during surgery by segmenting the image and generating instructions based on previous knowledge. Table 3.1 lists a few examples of the main applications of ViT in medical image segmentation.

Table 3.1: Examples of ViT application in medical image segmentation.

Application in Image Segmentation		
Method	**Category**	**Medical Application**
TransUNet [137]	MRI, CT	CT and MRI cardiac segmentation
Dual-TransUNet [138]	Microscopy	Skin lesion analysis [168]; Gland segmentation in histology [169]; Nuclei in divergent images [170]
Swin-Unet [158]	CT	Abdominal multi-organ segmentation
iSegFormer [160]	3D MRI	Knee image segmentation
3D UX-net [161]	3D MRI	Brain tissue segmentation
UNesT [162]	MRI, CT	Abdominal multi-organ segmentation + Kidney segmentation + Whole brain segmentation.
NestedFormer [163]	MRI	Brain tumor segmentation
RECIST [164]	CT	Automatic tumor segmentation and diameter size prediction
GT U-Net [165]	X-ray	Tooth therapy: Root canal segmentation
FCN-Transformer [166]	Colonoscopy	Colorectal cancer segmentation
TraSeTR [167]	Endoscopy	Robot-assisted surgery

3.3.2 Application of ViT in Medical Image Detection

Image detection plays a crucial role in digital health and imaging analysis to identify objects within complex structures and share the information within the healthcare information system for further analysis. This is important to measure cell size and count the number of suspicious objects or malicious tissues.

Object detection is essential in cancer assessment when it is difficult to label or classify the cells, and a careful analysis is required to identify cancers. Detection-Transformer (DETR) is proposed to detect Lymphoproliferative diseases in MRI T2 images [171]. The size of metastatic lymph nodes is small in the MRI scans, and it is difficult to be identified. The performance of the (DETR) can reduce the false positive and improve the precision by 65.41% and sensitivity by 91.66%.

The Convolutional Transformer (COTR) [172] detects polyp lesions to diagnose Colorectal Cancer (CRC) using colonoscopy images. CRC has the second mortality danger in the world, among other cancers. COTR architecture employs CNN for feature extraction and convergence acceleration. At the same time, the Transformer encoder is used to encode and re-calibrate features. On the other hand, the Transformer decoder is utilized for object querying, and the feed-forward network is for object detection.

Global lesion detection in CT scans is detected by Slice Attention Transformer (SATr) [173]. The backbone of (SATr) is composed of a hybrid combination of convolution and Transformer attention to detect log-distance feature dependencies while preserving local features.

Lung nodules detection is investigated using an unsupervised contrastive learning-based Transformer (UCLT) [174]. Lung nodules are the small masses in the Lung, which is a form of cancer. It is small and difficult to be detected within the complex lung structure. The research harness the contrastive learning and vision transformer to break down the volume of CT images into small patches of non-overlapped cubes and extract the embedded features to process them using the Transformer attention mechanism.

In order to predict the hemorrhage category of Brain injury in CT scans, a Transformer-based architecture is also used for Intracranial Hemorrhage Detection (IHD) [175]. Table 3.2 summarizes the ViT applications of image classifications.

Table 3.2: Examples of ViT application in medical image detection.

Application in Image Detection		
Method	**Category**	**Medical Application**
DETR [171]	MRI	Lymphoproliferative diseases detection
COTR [172]	Colonoscopy	Colorectal cancer detection
SATr [173]	CT	Universal lesion detection
UCLT [174]	CT	Lung nodule detection
IHD [175]	CT	Brain injury hemorrhage detection

3.3.3 Application of ViT in Medical Image Classification

Classification is of the essential digital health solutions in medical imaging analysis to help the medical practitioner automatically identify the object within a complex structure to categorize the medical case immediately. Utilizing AI while working in remote areas and using telehealth systems with limited medical resources is critical to ensure the accuracy of final clinical decisions. The importance of AI emerged during the pandemic while there was pressure on the healthcare system that overloads the capacity of the healthcare infrastructures. There are different applications for the Vision-Transformer in medical imaging classification.

TransMed [4] uses the combination of Transformer and CNN to classify multimodal for medical analysis. The classification includes disease classification and lesion identification. Figure 3.10 depicts an example of using TransMed in image classification example.

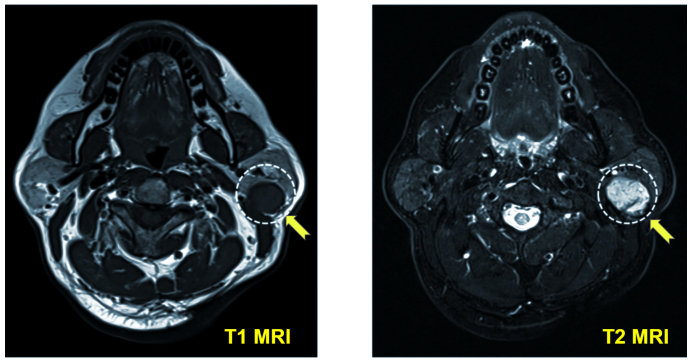


Figure 3.10: Example of using Vision-transformer for tumour classification in MRI images using TransMed [4]. Tumour is annotated by a yellow arrow and circle on the brain image.

Shoulder implant manufacture [176] uses Transformer in Orthopedics applications to assist in shoulder replacement surgery, including artificial implants and joints. The research uses shoulder X-ray images to detect and classify the implanted shoulder manufacturer before the surgery to identify the required accessory. GasHis-Transformer [177] proposes a multi-scale Visual-Transformer for detecting and classifying Gastric cancer images using histopathological images of the Hematoxylin and Eosin using a microscope. A comparative analysis of Cervical Cancer classification using different deep learning algorithms, including Transformer, has been investigated using cytopathological images [178]. Brain metastases classification [179] used a Transformer-based model to classify the whole brain MRI. Brain metastases are one of the causes of malignant tumors in the central nervous system [180]. ScoreNet [181] is an implementation of the Transformer model to classify Breast cancer using histopathology images. RadioTransformer [182] classifies COVID-19 cases based on chest X-ray images. TractoFormer [183] classifies Brain images based on Tractography which is a 3D modelling of the brain nerve tracts using diffusion MRI. TractoFormer discriminates 3D fibre spatial relation. They reported a decent accuracy in classifying schizophrenia vs control. Table 3.3 summarizes the ViT applications of image classifications.

Table 3.3: Examples of ViT application in medical image classification.

Application in Image Classification		
Method	**Category**	**Medical Application**
TransMed [4]	MRI	Multi-modal classification: disease classification, lesion identification
Shoulder Implant Manufacture [176]	X-ray	Orthopedics: Shoulder implant manufacture classification
GasHis-Transformer [177]	Histopathology Microscopic Images	Gastric cancer classification and detection
Multi-Scale Cytopathology [178]	Cytopathological Images	Cervical cancer classification
Brain Metastases Classification [179]	MRI	Classification of the brain tumor of central nervous system
ScoreNet [181]	Histology Datasets of Haematoxylin + Eosin	Breast cancer Classification
RadioTransformer [182]	X-ray	COVID-19 Classification using Chest X-ray Images.
TractoFormer [183]	Diffusion MRI	Nerve tracts modelling and 3D fiber representation

3.3.4 Application of ViT in Medical Imaging Prognosis Predication

The ability of Vision-Transformer to analyze time series sequence data and get insight from the previous data allows for predicting future behaviour or pattern. In medical imaging, it is essential to help the healthcare practitioner predict the effects of diseases or cancers to treat them before they spread more.

Fig. 3.11 depicts using the Transformer for surgical instructions that are also implemented in (SIGT). The algorithm uses Vision Transformer to analyze the vi-

sual scene during the surgery and update the reinforcement learning reward and state status to predict the instruction for the robot.

Figure 3.11: Example of using ViT for surgical instruction prediction. Transformer prediction is based on the SIGT method [5]. Ground truth is used as a reference for comparison and validation.

Sig-Former [184] can predict the surgical instructions during the operation using the Transformer attention mechanism to analyze the input image. The dataset includes images like laparoscopic sleeve gastrectomy and laparoscopic ventral hernia that are acquired during surgery.

3D Shuffle-Mixer (3D-SMx) [185] analysis 3D volumetric images from CT and MRI using context-aware dense prediction for different diseases, including hemorrhagic stroke, abdominal CT images, and brain tumor. The Graph-based Transformer models (GBT) [186] also predict genetic alteration. Ultrasound recordings are used for fetal weight prediction by the Residual Transformer Model (RTM) [187]. CLIMAT [188] forecasts the knee osteoarthritis trajectory based on the X-ray images from specialized radiologists. Table 3.4 lists the summary of some examples of using ViT for image prediction.

Table 3.4: Examples of ViT application in medical image predication.

Application in image prediction		
Method	**Category**	**Medical application**
3D-SMx [185]	3D (MRI,CT)	Context-aware dense prediction for different diseases that includes hemorrhagic stroke, abdominal CT images, Brain Tumor
GBT [186]	Cancer Genome (TCGA)	Computation Pathology: Genetic Alteration
RTM [187]	Ultrasound	Fetal weigh at birth prediction
CLIMAT [188]	X-ray	Forecasts Knee Osteoarthritis Trajectory
Sig-Former [184]	Laparoscopy	Surgical Instructions Prediction
SIGT [5]	Robot camera	Surgical instruction prediction and image captioning

3.3.5 Application of ViT in Image Reconstruction and Synthesis

After acquiring the data from the medical imaging modalities like MRI, CT, and Digital X-ray, the images are stored as raw data in an unstructured format. In order to make this raw data readable, a reconstruction process should be constructed to retrieve the image without any loss. This is computationally expensive due to the size and complexity of the reconstruction algorithms. Using Deep learning significantly impacted the reconstruction performance. It enhanced preserving the fine image details within a reconstruction time of a few seconds. It is needed more time in traditional techniques like image reconstruction using compressed sensing [189]. Reconstructing Magnetic Resonance Images is challenging due to the K-space matrix's size, complexity and sparsity, where the raw images are stored in the frequency domain.

SLATER [190] a Zero-shot adversarial Transformer is used to reconstruct MRI images using unsupervised training. SLATER maps the noise and latent representation to MR coil-combined images. To maximize the consistency of the im-

ages, operator input and the maximum optimized prior information is combined using a zero-shot reconstruction algorithm. Fig. 3.12 depicts different methods for reconstructing fastMRI and the reconstruction error map using SLATER (ViT-based method) from T_1 weighted images and compare them with other techniques that are based on non-ViT methods. (ZF) is the traditional Fourier method [191], LORKAS [192, 193], GAN$_{sub}$ [194], SSDU [195], GAN$_{prior}$ [196], and SAGAN [197] are Generative Adversarial Network reconstruction-based methods, SLATER is Vision-Transformer based-method [190].

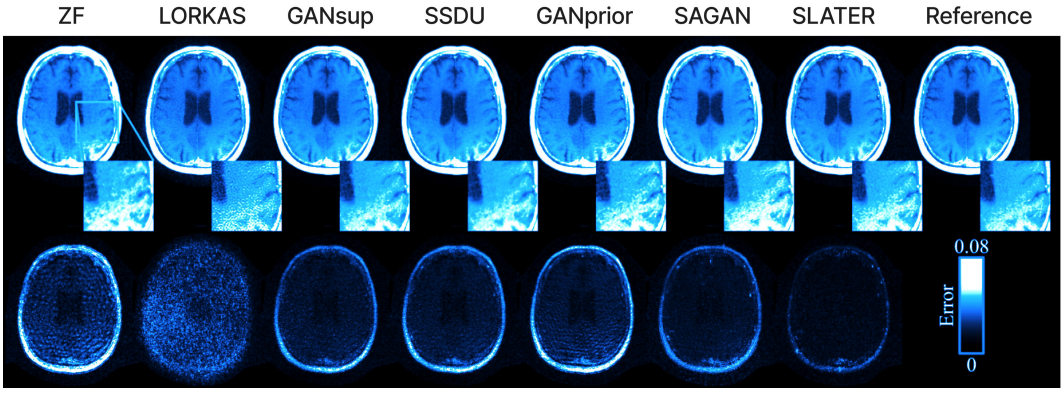


Figure 3.12: Top: Different reconstruction methods from T_1 weighted acquisition of the fastMRI using different methods. Bottom: reconstruction error map.

Task Transformer (T^2Net) [198] proposes an architecture to jointly reconstruct and image enhancement using a super-resolution method for MRI. The process of the T^2Net can be explained in two points. First uses two CNNS sub-tasks to extract domain-specific features. Second, T^2Net embedded and synthesized the relationship between the two sub-tasks. ReconFormer [199] addresses the problem of the under-sampled K-space data by utilizing the Recurrent Pyramid Transformer Layers to retrieve the data rapidly and efficiently. Utilizing Transformer-based methods for fast MRI reconstruction is evaluated in work [200]. The results show that the combination of Generative Adversarial Networks (GANs) and ViT got the best performance and is 30% better than the standard methods like Swin-Transformer. Table 3.5 summarize the ViT application in the image reconstruction.

Table 3.5: Examples of ViT application in medical image reconstruction.

Application in Image Reconstruction		
Method	**Category**	**Medical Application**
SLATER [190]	MRI	MRI Unsupervised Reconstruction
T^2Net [198]	MRI	Image Reconstruction and Super-Resolution Enhancement
ReconFormer [199] and FastMRIRecon [200]	MRI	Accelerated MRI Reconstruction
E-DSSR [201]	Endoscopy	Surgical Robot Scene Reconstruction
DuTrans [202], MIST-net [203]	CT	CT Sinograms Reconstruction

ViT (Stereo Transformer) is utilized by E-DSSR [201] to reconstruct the robotic surgery scene that is acquired by Endoscope. This application is essential in surgery education, robotic guidance and context-aware representation.

DuTrans [202] uses Swin-Transformer in the core of their architectural design to reconstruct the sinograms of the CT scans from the attenuation coefficient of the Hounsfield Unit [204]. Efficient reconstruction of CT scans is essential to get high-quality images, reduce radiation dose and distinguish fine details to help early detection of cancers.

MIST-net [203] proposes a multi-domain Transformer model to reconstruct CT scans. MIST-net can lower radiation doses without compromising image quality. MIST-net incorporates the Swin-Transformer Architecture, residual features and edge enhancement filter to reconstruct the desired CT image.

3.3.6 Application of ViT in Telehealth

There is an emerging need for efficient techniques to process all medical information within the healthcare ecosystem. This is because of the complex nature of the unstructured format of medical data, either an image, clinical report or laboratory results. ViT provides a comprehensive solution as it can process medical data in a different format and automatically generate reports or instructions. Fig.3.13

depicts the telehealth ecosystem's main components: data source, ingestion, machine learning and data analysis.

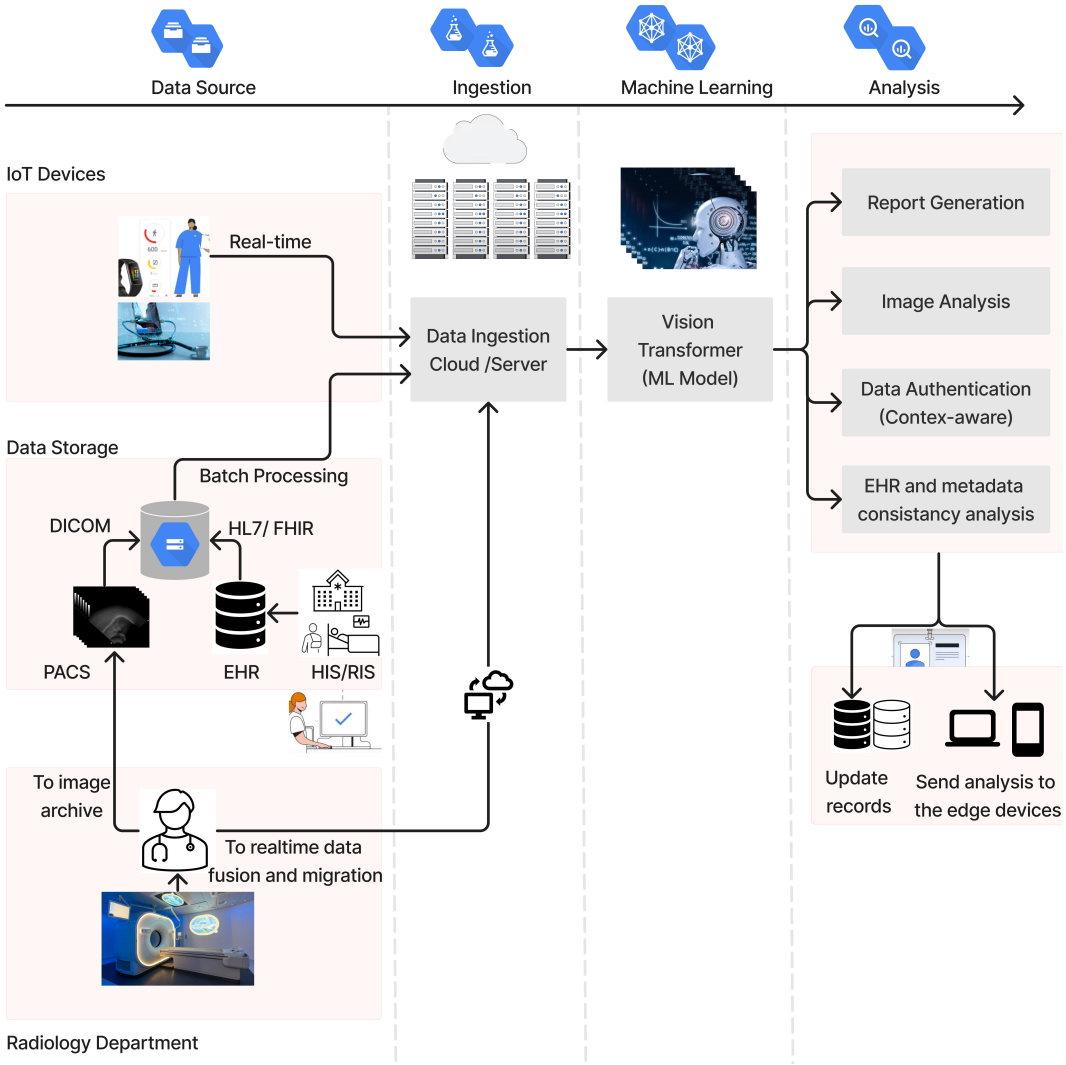


Figure 3.13: Illustration of using Vision-Transformer in telehealth ecosystem.

The Hospital Information System (HIS) and the Radiology Information System (RIS) register the patient and store the data in the EHR and PACS to be shared within the telehealth ecosystem. The hospital information system relies on the standards like HL7/FHIR [205, 206] to exchange patients metadata or EHR reports. While Picture Archive and Communication System (PACS) [207] is used to store and transfer medical images mainly in Digital Imaging and Communications in Medicine (DICOM) [208] format to be available to medical staff for further clinical

analysis.

Patient data are shared with the cloud or server either in real-time streaming or batches from the data storage warehouse or data lake in the case of working in the cloud environment. Vision-Transformer or any other machine learning model is used to train the system on the ingested data. Once the model has been deployed, ViT can be used to analyze medical data, which has about 90% as an image format and the rest for other types of data. Once the data is analyzed, the results are sent to update patient records in EHR or other storage systems.

Application of ViT in Report Generation

Vision-Transformer provides a unified solution to process text alongside unstructured data like images. The advantage of using ViT is to process and generate radiology reports, surgical instructions and other clinical reports in a global context by retrieving huge amounts of information stored in health information systems.

Fig. 3.14 illustrates image captaining, report consistency, completeness and report generation by (RTMIC) [209] and (IFCC) [210] from an input of medical images. RTMIC is a Vision Transformer-based algorithm that is used for medical image captioning (RTMIC) [209]. Ground Truth is a manual reference that is written by a professional expert. Att2in [211] is an attention-based method used for comparison. The quality standards of the health information system state that the transferred data should be consistent and complete. The work (IFCC) [210] improves factual completeness and consistency of image-to-text radiology report generation. The algorithm uses a hybrid combination of Transformer to extract features and reinforcement learning to optimize the results.

Transformer efficiently addresses the challenges of handling biased medical data and long and inconsistent paragraphs. AlignTransformer [212] can produce a long descriptive, and coherent paragraph based on the analysis of medical images. AlignTransformer works mainly in two stages. First, align the medical tags with the related medical image to extract the features. Second, use the extracted features to generate a long report based on the training data of each medical tag.

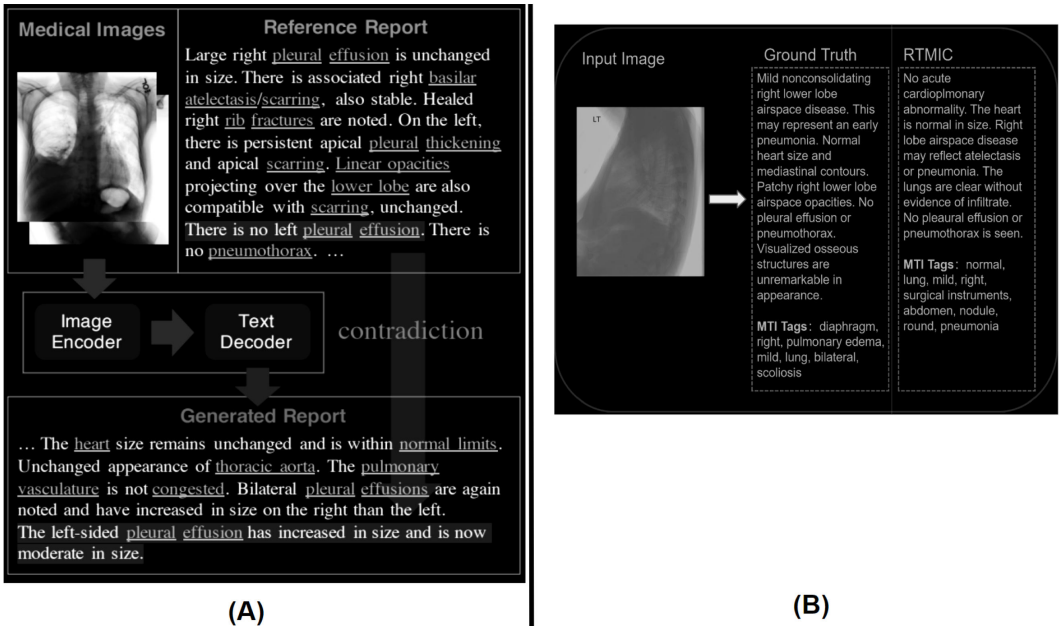


Figure 3.14: Examples of report generation from the input image using Vision-Transformer. (A) Sample of results by (IFCC) for report completeness and consistency. (B) Example of report generation results using (RTMIC).

Transformer is also used to generate surgical reports during the robot-assisted surgery by learning domain adaptation in (LDASR) [213]. LDASR uses a Transformer to learn the relationship between the desired region of interest, surgical instrument and image to generate image captioning and report during surgery. Table 3.6 summarizes the ViT application in image generation.

Table 3.6: Examples of ViT application in medical report generation.

Application in Medical Report Generation		
Method	**Category**	**Medical Application**
RTMIC [209]	Medical images general	Report generation from medical Images (e.g. MRI, CT, PET and X-ray)
IFCC [210]	Medical images general	Medical report completeness and consistency
AlignTransformer [212]	Medical images general	Long report generation from medical images tags
LDASR [213]	Surgical Robot Camera	Surgical report generation

Possible Application of ViT in Telehealth Security

Telehealth security is getting significant attention from healthcare providers due to the emerging risk associated with leveraging advanced AI-based technology. Generative AI may poses serious risk if the process of collecting and processing data are not managed properly. In healthcare, there is a serious risk of misdiagnosing a patient with the wrong disease or even diagnosing a healthy person with a disease. User-based error either the user is human or robot occupy the majority of the error in the healthcare information system.

The adversarial attack is the common phrase for malicious attempts of the machine learning algorithm or data vulnerability. Attacks may include modifying the data or the algorithm code, causing a wrong output [214, 215]. The algorithm's accuracy may also be affected as the code, model weight or labeled data could be manipulated. Cybercriminals aim to threaten healthcare providers to get some money. Otherwise, they will publish the patient's information to the public, encrypt the database, and not give them the decryption key.

Fig. 3.15 shows an example of the effect of data poisoning by adversarial attacks in restoring the original image after malicious perturbation during the machine learning model training.

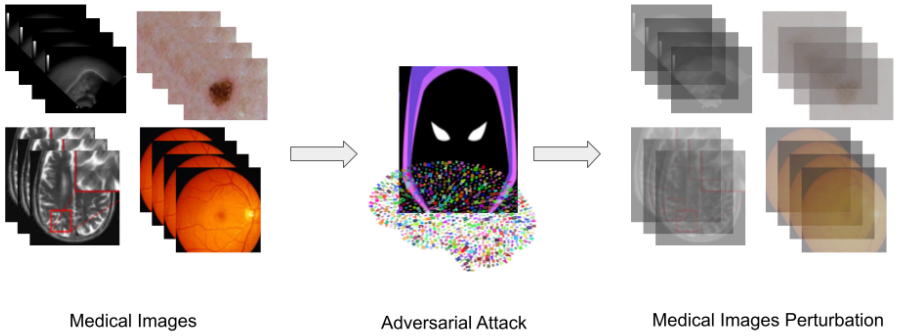


Figure 3.15: Illustration of data poisoning by the adversarial attack to fool learning-based models trained on medical image datasets.

Listed below are some examples of countermeasures used to combat cybercrimes that developed by the industry.

1. Implement a context-aware system to ensure that the implemented code is safe and not jeopardized.

2. Store data in an encrypted cloud environment to guarantee that the data is safe and backed up.

3. Federated learning is another measure that uses a distributed computing engine to process the data in a geographically distributed environment that keeps the data in different locations and makes it difficult for hacking.

4. Embrace zero-trust policy to manage access control systems in digital health applications. This will add additional authentication measures by considering different attributes to grant access permissions and not just relying on a role-based access system.

The traditional CNN's-based algorithms are not robust against adversarial attacks due to the simple nature of the CNNs architecture to make it vulnerable to any malicious attempts. On the other hand, Vision-Transformer is more robust to adversarial attacks compared to CNN's [216]. The complexity of the ViT algorithm and the ability to extract features in a global context is a solid ground to detect any irregularities in data entry. ViT is used for data encryption [217], anomaly detection [218], network intrusion system detection [219], anti-spoofing [220] and patch

processing [221] for distributed data storing of. Table 3.7 summarizes the ViT application in the image reconstruction.

Table 3.7: Examples of ViT Application in Security .

Application in Security	
Method	Application
Jigsaw Block-based Encryption [217]	Data Encryption
MFVT [218]	Anomaly Detection
Image Conversion from Network Data-Flow [219]	Network Intrusion System Detection
Zero-Shot Face [220]	Anti-spoofing
Backdoor Defender [221]	Patch Processing

3.4 Limitation and Challenges of ViT in Digital Health

The Transformer-based algorithm is emerging as the state-of-art in vision tasks to replace the traditional standalone-CNN architectures. However, Transformer-like any ML model, has limitations and challenges that are either technical or regulatory compliance requirements. The primary limitations include data size and labelling, the need for a hybrid model, data bias and model fairness, and ethical and privacy challenges.

3.4.1 Dataset Size and Labeling Challenges

Like any other attention-based mechanism, Transformer inherently requires a huge amount of data to train the model. Transformer achieved the best performance compared to the well-known ResNet when trained on the JFT dataset [222] that contains 300 Million images and 18 thousand classes, while its performance is compared to it when trained on ImageNet-21k [223], that have about 14 Million images and 21 thousand classes. On the other hand, the Transformer performance does not surpass the ResNet architecture when trained on ImageNet-1k [224, 225], which has 1.28 million image examples and one thousand classes.

Fig. 3.16 visualize the performance of different ViT and ResNet architectures with respect to the data size. The results show that ResNet has better performance

when the dataset is small. ResNet and ViT have almost the same performance when the training is about 100 Million samples. However, ViT gets superior performance compared to ResNet when the dataset size is larger than 100 Million images.

The limited dataset size is challenging in medical applications as it is difficult to find a clean and high-quality dataset that is feasible for clinical application standards. Moreover, finding qualified specialists to annotate millions of images is not easy, expensive and time-consuming.

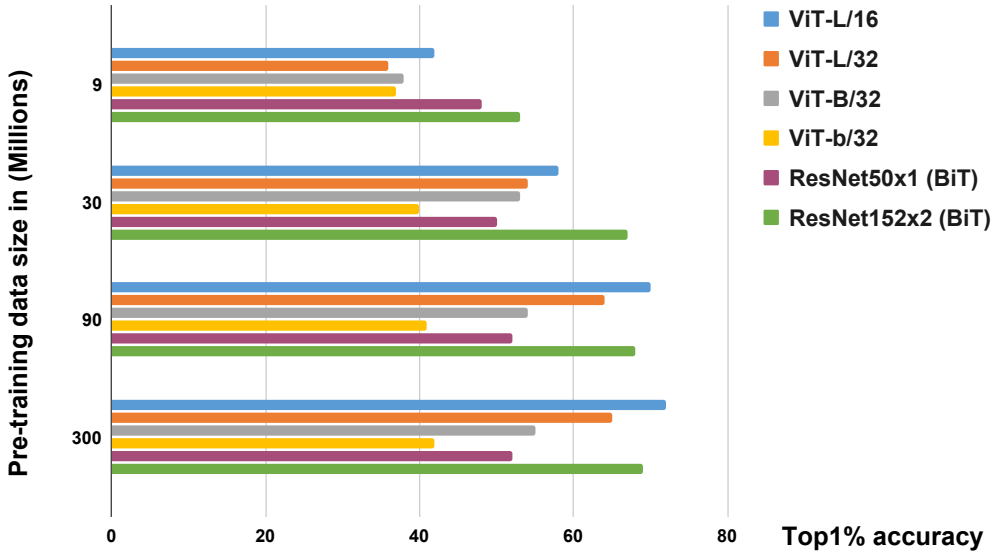


Figure 3.16: Comparison between different Vision Transformer (ViT) and ResNet (BiT) architectures accuracy to the size of different subsets of training data. Y-axis is the data size of pre-training in the ImageNet dataset. The X-axis is the accuracy that is selected from the top1% of the selected 5-shots of ImageNet. The results according to the study in [2].

Transfer learning, data augmentation, adversarial imaging synthesis and automatic data labelling could be best practices to alleviate the insufficient dataset size. However, for the best performance of the ViT model, the researchers in [226] suggest that it outperforms ResNet when trained from scratch on the large Imagenet dataset without using data augmentation or a large pre-trained model. It is a trade-

off between dataset size limitations and performance, as having a large dataset and enough computational resources to train them is challenging. Using cloud-based data training can be a solution for limited resources. However, it is a more expensive option for academia and more applicable for industrial applications. In addition, the study in [227] suggests an effective weight initialization scheme for fine-tuning the Vision-Transformer using self-supervised inductive biases learned directly from small-scale datasets.

Contrastive Learning (CL) is also being used in conjunction with Transformer. CL is essential in medical image applications as it can help to minimize the difference between similar object representations in the latent space while maximizing the difference between the non-similar objects [9]. CL has been used with Transformer in medical histopathology when it is needed to classify huge size images (in Gigapixels) and get inference to distinguish between multi-label cancer cells to classify them [228].

3.4.2 The Need for Hybrid Model with Transformer

Transformer is initially designed to process language models in a sequence format, and then it has been modified to process vision tasks by splitting the image into small patches and processing them sequentially as a text-like model. However, Transformer can get inferences about the information in a global context to capture a wide range of dependencies between objects, but it has a limitation in features localization. While Transformer-stand alone model is sufficient for most of the classification tasks but in the case of image segmentation for critical medical applications, when a high-quality image is needed, Transformer performance is not enough, and it needs to be combined with a hybrid model.

UNet or ResNet architectures are widely used as standard models for medical image segmentation that can preserve image details due to the nature of encoder-decoder architecture with residual connections. However, UNet and ResNet inherit the limitation of CNN for not capturing a wide range of dependencies by having local feature extraction capabilities only. TransUNet [137] is the first architecture proposed for medical imaging segmentation that combines Transformer and Unet architectures for local and global feature extraction.

The Transformer is also combined with Reinforcement Learning (RL) to generate instructions for surgical robots [5, 184]. RL is essential while working in

a robot environment. The Transformer can capture features to update the state-reward status in RL to automate robot tasks. RL-Transformer combination is also used in medical image captioning [209] to generate medical reports automatically within the hospital system.

3.4.3 Data Bias and Fairness

Training machine learning models using huge datasets (In millions or billions of examples) requires suitable infrastructure with enough computational and storage resources. For this reason, many algorithms tend to apply dimensionality reduction to minimize the model parameters, which reduces the extracted features. This allows us to train the model using less computational or memory requirements. However, there is a chance of losing some information that has less representation in the feature map or dataset. Consequently, the model might be biased toward the labels or classes with the largest training data. The bias in the results can be significant, especially when the label balancing is not performed before training. In the case of medical applications, rare diseases and any outliers could be disregarded from the model prediction.

In the study, [6], the fairness and interpretability of the Deep-Learning (DL) models are evaluated using the largest publicly available dataset MIMIC-IV (Medical Information Mart for Intensive Care, version IV). The study finds that some DL models lack fairness when relying on demographics and ethnicity to predict mortality rates. On the other hand, the DL models that use proper and balanced critical features to train the models are not biased and tend to be fair. In many models, racial attributes are used unequally across subgroups. This causes inconsistent recommendations for using mechanical ventilation for treatments or not in the intensive care unit when relying on demographic and racial categories like gender, marital status, age, insurance type and ethnicity. Fig. 3.17 shows examples of global futures importance ranks that are used to predict the mortality rate using different ML methods [6]. The figures depict how the importance score is biased for some features when the ML algorithms are changed.

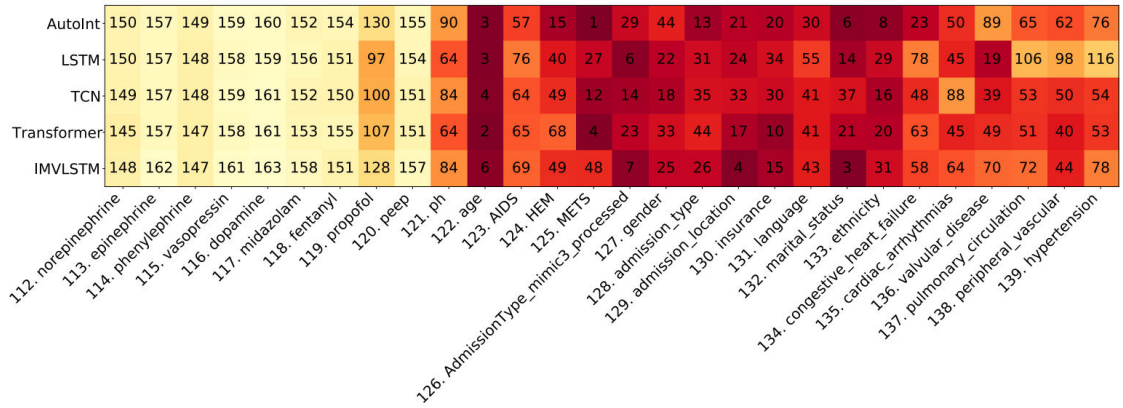


Figure 3.17: Visualization of heat map for the comparison between different examples of global futures importance ranks that are used to predict the mortality rate using different ML methods [6].

3.4.4 Ethical and Privacy Challenges

Sharing information in the healthcare information system is regulated. Privacy and ethical regulations might be different from one jurisdiction to another. For example, The Health Insurance Portability and Accountability Act (HIPAA) regulates the healthcare information system. HIPAA is a national standard in the United States of America (USA) to protect sensitive patient information. The standard states that the information can not be disclosed without patient consent. At the same time, patients have the right to access their data, ask for modification and see who accessed them. While the regulations help preserve patient privacy, collecting health-related detests or making them available to the public is challenging. In the case of the Vision-Transformer, this is critical as millions of examples are needed to train the model to get accurate results.

Using a Vision-Transformer or any other machine learning model that is trained on a large detest is error-prone, and the results are subject to ethical concerns. Many large datasets are scrapped from the internet as the source may be unknown or untrusted, and there is no previous consent to collect these data. Training the Vision Transformer from untrusted sources can generate false results that could cause an error in the automatically generated report for the patient. Offensive con-

tent can also be generated in the medical report. The consequence might be worse in the case of data breaches or cyber-attacks on the healthcare information system that may alter the patient records, images or the performance of the data streaming of the telehealth system.

Vision Transformer is more robust against adversarial attacks, but there is no guarantee that the ViT-based model can not produce inappropriate content. This raises a red flag to regulate the current AI industry and the overall process when it is used for healthcare to ensure the input and output of the systems are clean and valid for clinical applications. Federated learning from different healthcare facilities and different edge devices or servers can help maintain data privacy at a high level. However, the research in [229] reported a vulnerability of retrieving the original data even if you only have the wight or learned model.

Chapter 4

Effective Tongue Contour Segmentation Using Well-Managed Dataset and ViT-based Architecture

This chapter discusses the process of utilizing TongueTransUNet for object segmentation. The chapter highlights the process for implementing typical machine learning model. The architecture design of TongueTransUNet. In addition, the feature extraction and embedding strategy in latent space were highlighted. The chapter also presents the quality control (QC) process for using a well-managed process to achieve the most desired results. The QC considers different measures like adaptive and normalized accuracy and add an additional human-interactive reinforcement feedback to correct the automated model.

4.1 Machine Learning Model Pipeline

This section contains three main items. First, it presents the road map for implementing Vision Transformer and any other typical ML model. Second, the architecture design for the proposed TongueTransUNet is used to segment the tongue using Vision Transformer. Third, the design for a zero-trust context-aware system will ensure secure communication between the ultrasound system and other medical devices with the cloud-based health information system.

4.1.1 Road Map for Implementing Typical ViT-based Model

Fig. 4.1 depicts the four main steps to implement the ViT model pipeline from end to end. The first stage is problem formulation. The second is data processing. Third is model implementation, training and validation. Forth model deployment and quality assurance.

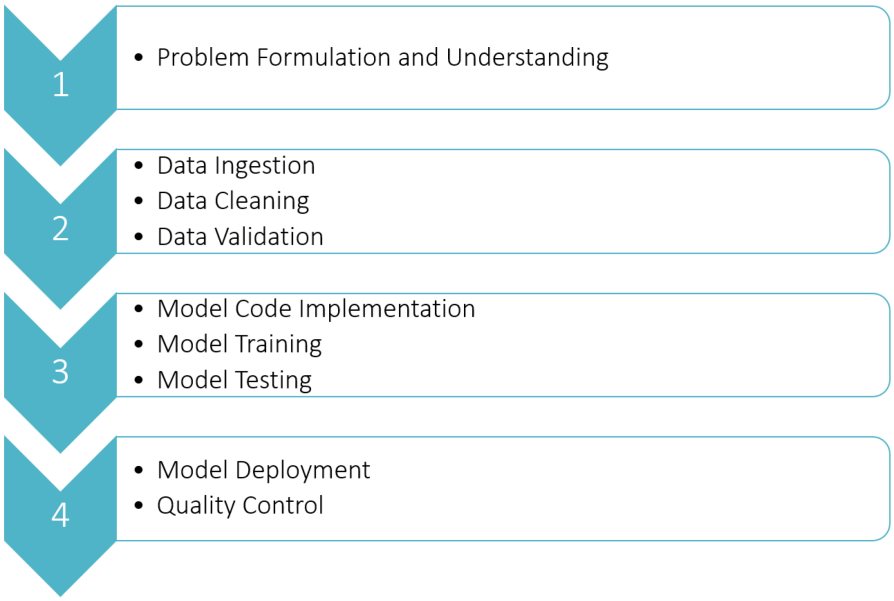


Figure 4.1: Road map for vision transformer implementation.

Problem Formulation: The first step before implementing any design is understanding the problem and formulating it to fit within the context of the desired product use case.

Data Preparation: Once the problem is understood, preparing high-quality data is the essential step in any Artificial Intelligence (AI) algorithm. The data should be relevant, correct, statistically balanced and sufficient for training in case of different classes are used. The data should be verified by using different qualitative and qualitative measures to guarantee that the ingested data by the AI model are valid for training. This will help the model to be stable during the training and converge faster to reach the optimal solution. During the

Model and Code Implementation: There is no master algorithm that fits everything, and each one has advantages and disadvantages. The suitable ViT model

or architecture is selected based on the available data and problem use case to get the desired success metric. The model hyper-parameters are fine-tuned during the training stage to reach the desired accuracy and prevent model over or under-fitting. The model is also validated and tested on separate datasets other than the data trained on.

Model Deployment and Testing: Finally, when the model passes all end-to-end testing and verification processes, it should be ready for deployment. There are different environments to deploy the final product on different applications, either mobile apps, web applications or on-premises software. The recommended environment is the cloud-based system, as it can automatically generate the model on a scale to fit the computation resources for the different applications. The deployed model should go through different quality assurance and monitoring to ensure that the targeted performance of the system is achieved during the test outside the lab and fix any bugs in the code or train the system on new data.

4.2 Experiment Dataset Information

This section discusses and highlights the data sources used in this research. There are two main sources: first, data from Transfer Learning, which includes millions of additional samples. The second source is data augmentation, which adds versatility and variation to image analysis by utilizing the geometrical transformation of the still image. The images have been de-identified to create a valid and legal dataset for model training without violating personal information. The images were obtained randomly and do not contain specific signatures for any individual subject. It has been used only for feature engineering to identify image structure in computer vision tasks. The largest dataset used to train the model is from the transfer learning source. Then, the model is fine-tuned using the dataset annotated by an expert to be utilized as features in a latent space without using the original image context.

The non-ultrasound images used to test the model performance and initialize weight for the UNet architecture are from ImageNet [224]. Different images classless from ImageNet were selected randomly. The selected dataset is balanced to ensure each class has the same size.

4.3 TongueTransUNet Architecture and Methodology

The overview of the TongueTransUNet architecture and methodology flowchart is depicted in Fig. 4.2. The methodology process is explained in a series of processes that start from image ingestion, feature extraction, feature mapping, quality control, image reconstruction, and segmentation. Below the figure is a detailed explanation of the steps in the following subsections.

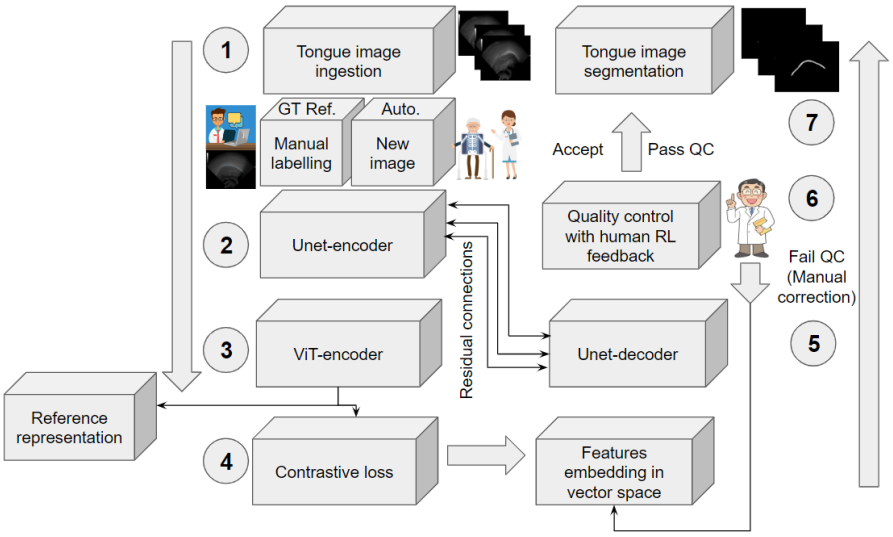


Figure 4.2: Illustration of the methodology overview. The process is annotated in a series of numbers from 1-7. GT Ref, is the ground truth reference that is manually annotated, and auto is the automatic processing for tongue ultrasound images.

4.3.1 Image ingestion

The input is ingested in the form of ultrasound images that were recorded using ultrasound. Fig. 4.3, depicts the process of ultrasound acquisition from the tongue. The ultrasound probe is placed below the chin to record ultrasound images of the tongue during speech. The tongue contour appears as a white arc. The tongue contour geometrical displacement is varied during the speech from different sounds or vowels to vowels.

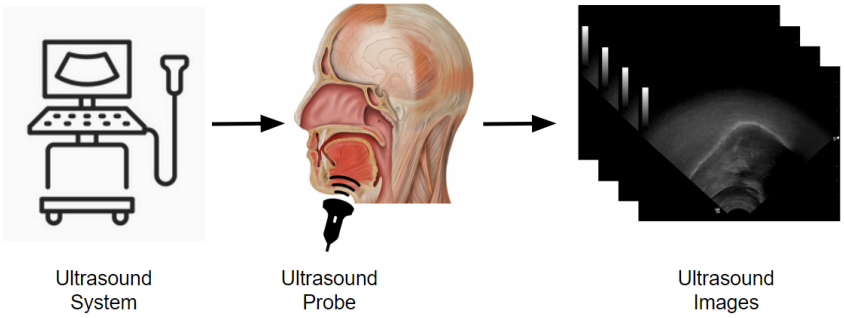


Figure 4.3: Input image using ultrasound system. Left: Ultrasound system. Middle: Head-transducer arrangement. The head and oral cavity picture was modified from the original picture for the case, courtesy of Associate Professor Frank Gaillard, Radiopaedia.org, rID: 35836, [1]. Right: Output of recorded ultrasound image, ultrasound image source [7].

The ingested ultrasound images are divided into two main subsets to process them in two different streams:

1. Manual processing (Ground truth): The proposed techniques rely on building a reference vector embedding using manually annotated images. Before the automatic processing, it requires a domain expert to annotate reference images and encode them in embedding space.

2. Automatic processing: The automatic tongue segmentation process uses the reference representation to compare the embedding of the automatically ingested images and decide if they belong to the ultrasound representation or have a different representation that does not belong to the image category. This step is important to have training data from related and high-quality images only and reject noisy or unrelated images.

4.3.2 UNet-based encoder

In the first stage of feature extraction, ultrasound image features are extracted using the UNet-based encoder, [137]. A UNet encoder is a stack of convolutional neural networks (CNNs) used to encode features deeply and extract fine image details. UNet-encoder is depicted in Fig.4.4 illustrating the stack of CNN layers

to extract the bottleneck features. Those features resemble the image details that will be used to identify and segment the tongue contour based on the labelled data. Residual connections between the UNet encoder and decoder are essential to preserve the image features from the first layers that could be filtered out in the deep architecture. The skip connection is also essential to prevent over-fitting and model bias. The extracted features from the UNet encoder are projected in a linear layer or bottleneck features layer before forwarding the features to the vision transformer layer for further processing and feature extractions, see Sec. 4.3.3.

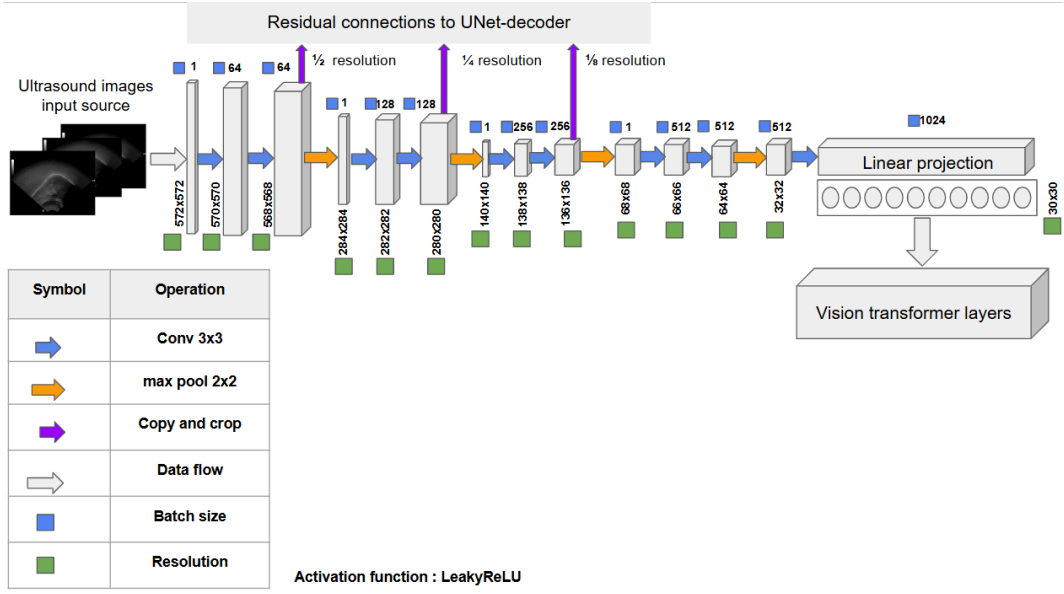


Figure 4.4: Representation of the UNet-based encoder diagram before the vision-transformer layer. The diagram depicts the batch size, resolution, and operations. Skip-connection link encoder and decoder through different resolution scales.

4.3.3 Vision transformer encoder layer

The extracted features by the UNet-encoder resemble the fine image details that help to segment the tongue contour. To further enhance the segmentation process by the UNet, a ViT layer is added to gain additional inference on the relationship and representation of the features in a wide perspective by using attention mechanism [135]. Fig. 4.5, depicts the process of parsing UNet-encoder features into the transformer layer. The process is initiated by splitting the features into a subset of patches and flattening them to fit the nature of ViT architecture. ViT maps the

features by extracting the embedding and position encoding. To further process the images, features are then normalized along the layer for better optimization. Multi-head self-attention is used to extract different characteristics in each head and score each feature to accept only the highest probability features [11]; in this research, we used 12 heads. Multi-layer perception is a type of feed-forward neural network with input, hidden, and output layers. The output layer contains the set of extracted features that has the highest probability or attention score. These features are used in this article to define the reference representation and lingual ultrasound image features.

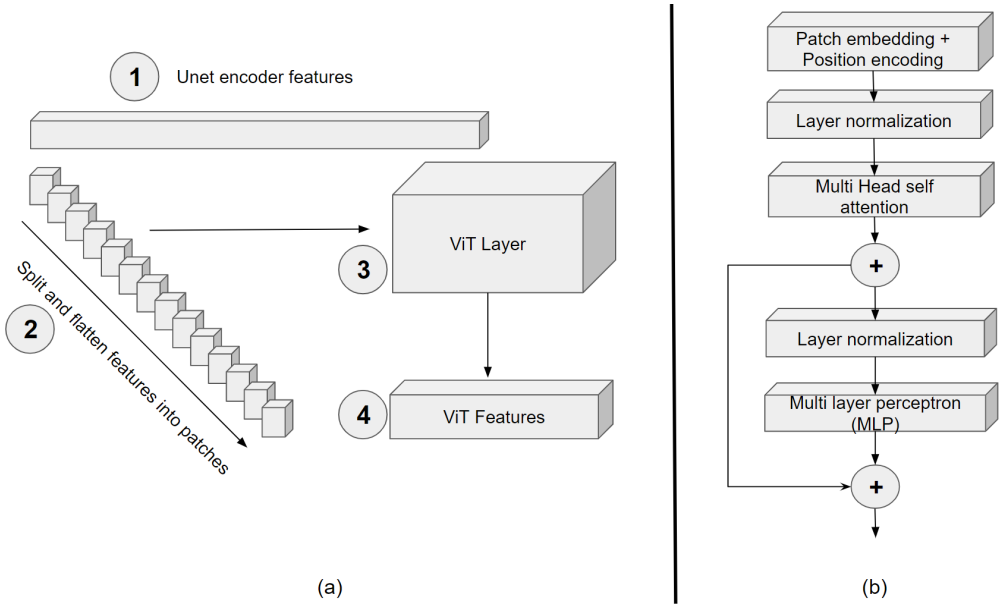


Figure 4.5: Representative of the vision transformer layers. (a) UNet features patch splitting and flattening before feeding them into the ViT layer. (b) Vision transformer layer. The sequence numbers represent the chronological order of the process.

4.3.4 Contrastive loss and feature embedding strategy

The feature embedding strategy utilizes human-interactive process alongside with automated data ingestion in embedding space. The combination is essential to add a trusted verification methodology within the overall process and to manage the quality of the both model input and output.

Human-based interaction utilized in two stages. In the initial stage human

expert required to create solid reference representation in the embedding space before the automated precess starts. At the later stages the human also added in the quality control stage as a reinforcement step if the automated process does not pass pre-defined threshold.

The automated strategy assign each feature or ingested input to the nearest feature in the embedding-reference representation of feature clusters for each related K-nearest neighbor. Fig. 4.6 shows the graphical abstract of the high-level logical flow of the feature embedding strategy.

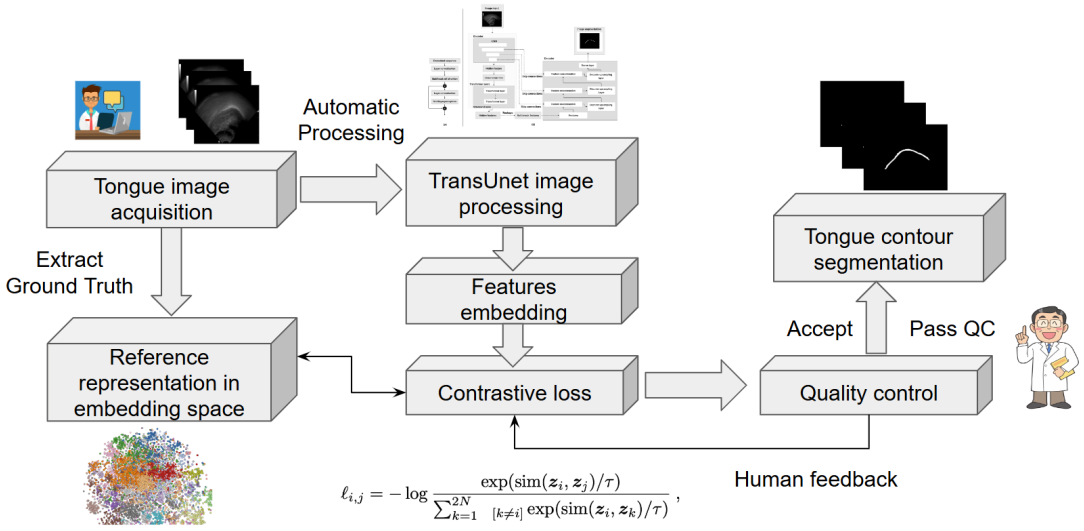


Figure 4.6: Graphical abstract of TongueTransUNet logical flow of the features embedding strategy.

Reference representation and contrastive loss (CL)

Contrastive loss is an essential component of AI data-centric approaches. It is efficient to map the encoded features in the embedding space to the right position. CL works by maximizing the agreement between similar features to make them closer to each other; this is noted as positive samples. At the same time, the dissimilar features are mapped to be apart from each other and noted as negative samples. The contrastive loss function used in this research is the *SimCLR*, Eq. (4.1) used in this research is inspired by *SimCLR* [9], see Fig A.1 in Appendix A.1 for more details.

$$l_{i,j} = -\log\left(\frac{\exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k\neq i]}\exp\left(sim(z_i, z_k)/\tau\right)}\right) \tag{4.1}$$

Where z_i, z_j are the two input images. $Sim()$ is the cosine similarity, and τ is the scaling factor. $1_{[k\neq i]}$, is an indicator function that denotes that it equals one if $[k \neq i]$; otherwise, it equals zero.

The process of using *SimCLR* in the proposed research is explained in the following step:

- Obtain the reference embedding by training a randomly selected 300 sample, which is annotated manually to serve as the ground truth embedding.
- Apply contrastive loss for any newly ingested image to compare it with the reference embedding.
- Reject any image that has embedding below the pre-defined threshold (Th_{CL}). Otherwise, accept similar images above the Th_{CL}.
- For any of the rejected images, apply manual annotation for the tongue contour. This helps to maximize the accuracy of the model to meet the medical standards.
- Continue the training for all samples and apply all previous steps.

Feature embedding strategy

Once the feature representation of each image is assessed by the contrastive loss assessment compared to ground truth features, the model creates a features cluster for the highest probability features.

Each cluster contains features with similar embedding. The cluster serves as a reference to guarantee that the related features are stored in a related cluster and far from other different features in other categories. The challenge of this technique is that it needs a human domain expert, and then the algorithm automatically assigns the new feature to the nearest cluster. Below is the explanation and hierarchical summary to the Fig 4.7 of the feature embedding process:

- The encoder extracted features and represented them in embedding space.

- The extracted features compared to the reference representation using contrastive loss, see 4.3.4.
- Derive the cluster center point for each K-nearest neighbor feature of each specific domain. In the case of this article, the focus is on tongue ultrasound images.
- Assign each feature to the nearest cluster of the K-nearest neighbor features in embedding space.
- Create a new cluster for each new feature that is not close to any pre-trained or manually annotated features.
- Update the embedding space in an iterative and incremental process during the training and quality control cycle. The embedded features in vector space are then decoded by the decoder to further process the image; see Sec. 4.3.5.

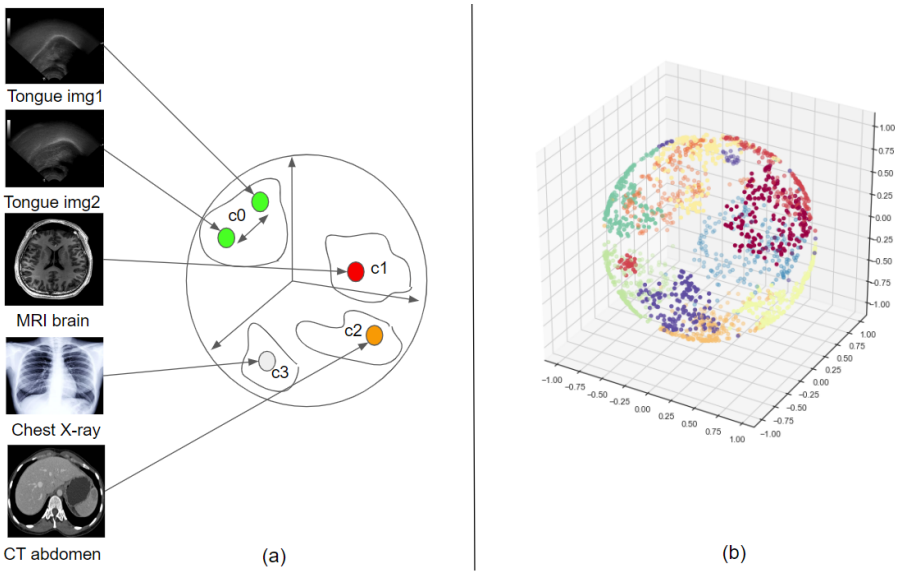


Figure 4.7: Illustration of features clustering at different indices at embedding space. (a) Visualization of vector space with a set of different features, where c(n) is the cluster center of each K-nearest neighbor of domain-specific features. (b) Representation of the features embedding in vector space.

4.3.5 UNet-based decoder

At this step, the processed features from the ViT are reconstructed using the UNet-based decoder to restore the full-size image. The decoder also receives input through residual connections directly from the encoder in step two and concatenates them to the ViT output. This is important to preserve features at the early stages that may be disregarded through the deep stack of filtering layers.

Fig. 4.8 depicts the TongueTransUNet decoder. The decoder architecture was modified from the UNet architecture to add a quality control process. The QC is implemented to filter the ingested data flow from the encoder to process only the high-quality images related to the tongue contour. The decoder stores image data in the vector space of the mapped features of MLP in the ViT layer. At the final stage in the decoder, the image is sent to the QC for final assessment before presenting it to the end-user or specialist, see 4.3.6.

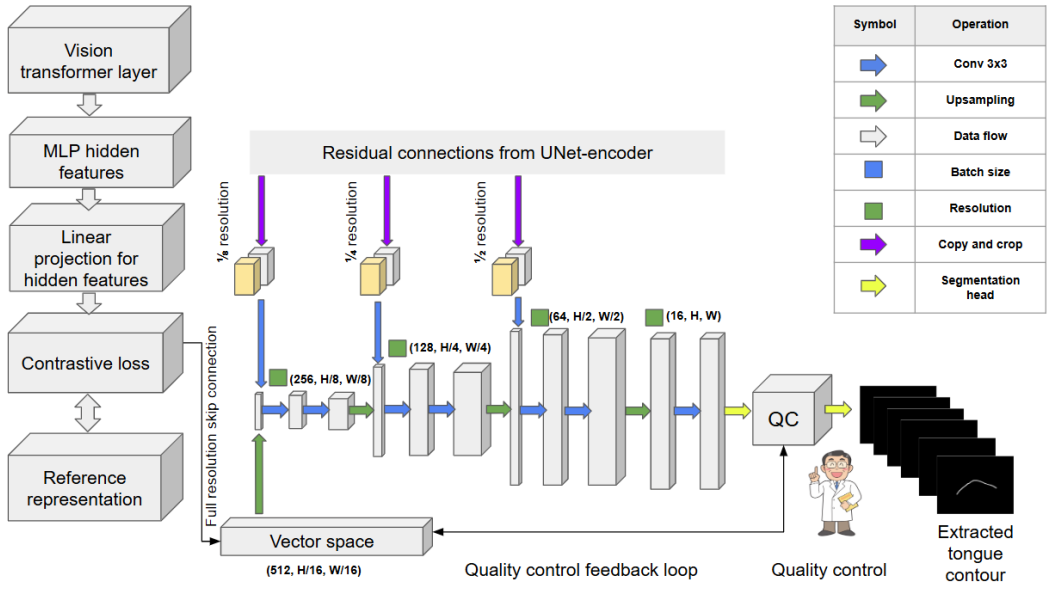


Figure 4.8: Visualization of the decoder architecture. Each block represents a feature that is used to reconstruct the image to its final shape. The resolution and operations of each block are described in the figure. H is the height, and W is the width. The batch size is similar to the encoder patch size.

4.3.6 Quality control (QC)

This section proposes a technique for controlling the quality of data training using a set of quality measures. First contrastive loss, Sec. 4.3.4. Second, the mean sum of distances for tongue segmentation accuracy, Sec. 4.3.6. Third, shape consistency, Sec. 4.3.6. Fourth, manual interactive segmentation, Sec. 4.3.6. Algorithm. 1 shows the hierarchical process for the QC during either image training or real-time segmentation.

Algorithm 1 An algorithm for tongue contour quality control (QC)

Input: Image (Img), Accuracy (MSD), Shape consistency (SC) , Contrastive loss (CL)

Set thresholds: Accuracy Threshold (Th_{MSD}), Shape consistency threshold (Th_{SC}) , Contrastive loss threshold (Th_{CL})

While input image = 1, otherwise verify input source
if $MSD \geq Th_{MSD}, SC \geq Th_{SC}, CL \geq Th_{CL}$ **then**
    Accept image representation.
    Update embedding space.
    Release tongue segmentation results.
else if $MSD < Th_{MSD}, SC < Th_{SC}, CL < Th_{CL}$ **then**
    Reject image representation.
    Reject tongue segmentation results.
    Ask for manual expert verification.
    Update embedding space.
    Release tongue segmentation results.
end if
    Output: MSD, SC, CL, Img

This process is not just important to improve the accuracy and quality of the final output, but it also helps to achieve better model convergence stability using smaller datasets. The standalone ViT requires millions of images to achieve high-quality results that are comparable to the well-known ResNet architecture. The TongueTransUNet can produce acceptable results when trained on a few hundred images, and it achieves clinically satisfactory results when trained in a dataset in a range of (10k-50k) images, where it was initially started with a few hundred manually annotated reference images. The reason for this large data range is that the model is dynamic and the quality performance depends on the quality of the

dataset. This increases the generalized performance of the segmentation results so the model can be sturdy and generalized for different data sources.

While model accuracy is important, obtaining high-quality results in smaller datasets consumes less computational resources, which minimizes the cost and time. However, the accuracy and computational complexity are not the only measures. Having clinically acceptable results is vital in the field of medical imaging analysis. The QC process ensures the model is generalized, accurate, and clinically valid by enforcing stringent QC processes that reject any unrelated or bad-quality images that are produced by the model due to noise, artifacts, or bad quality of the recorded input source.

Mean sum of distances (MSD)

Measuring the accuracy of segmented tongue contour is challenging as the contour length compared to the ground truth varies from frame to frame. This is due to motion artifacts, missing data, and dynamic movement of the tongue, changing the concavity of the tongue. To address this issue, the mean sum of distances measure is used as an evaluation measure for tongue tracking and segmentation; it was proposed by [20].

The MSD (mean square displacement) is a useful method for comparing the length of two contours. Other comparison methods, such as a mean sum of errors and norm, are not appropriate because the length of the two contours cannot be similar in most of the cases. The MSD is measured in pixels and can be converted to millimeters by assuming that each pixel is equal to 0.295 mm [15, 8]. Equation (6.1) shows the formula for the MSD.

$$MSD(U,V) = \frac{1}{m+n}(\sum_{i=1}^{n} min_j(\mid v_j - u_i \mid) + \sum_{j=1}^{m} min_i(\mid u_i - v_j \mid)) \qquad (4.2)$$

where (n) is the ground truth length and (m) is the extracted contour length, while (v_j) is the ground truth data points, and (u_i) is the set of extracted contour data points. While, (min_i) and (min_j) represent the nearest distances between each point on the contour and the nearest point on the other contour, respectively.

Shape consistency and area under contour

Measuring extracted contour accuracy or MSD is not enough to judge the quality of extracted color. MSD measures the normalized difference between the extracted contour and the ground truth. However, the shape consistency is essential to be assessed to make sure that the extracted contour has the shape of the tongue and is acceptable. Fig. 4.9 displays the tongue in an orange-dotted arc in the form of a semicircle.

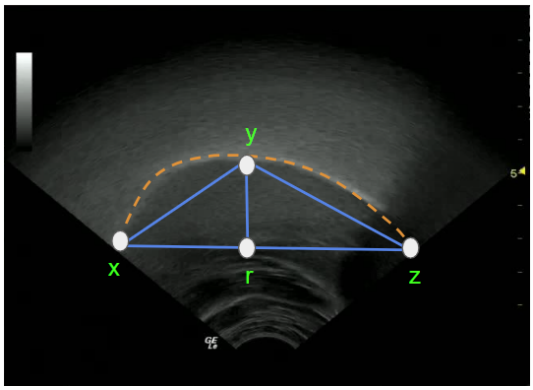


Figure 4.9: Shape consistency and are under tongue contour geometry visualization. The tongue contour is displayed in an orange-dotted arc. The points form a triangle head within the tongue contour semicircle.

The triangle in blue is used to assess the tongue shape consistency, and the semicircle is used to measure the area under the tongue. Equation (4.3) derives the area under the tongue. Equation (6.2) is used to measure the tongue contour curvature, while Equation (6.3) characterizes the asymmetry of the tongue contour.

$$A = \pi(xz)/2 \tag{4.3}$$

$$C = \frac{||yr||}{||xz||} \tag{4.4}$$

$$V = \frac{||xr||}{||rz||} \tag{4.5}$$

Where x, z, r, and y are the points on the tongue contour arc edges and the triangle in the area under the tongue contour.

Human-based Interactive Segmentation

If the confidence score is lower or the segmentation does not pass the QC threshold, the segmentation output is rejected. Additional measures must be added to verify whether the results should be disregarded or corrected manually. The algorithm prompts a screen with a list of images with a low QC score to be verified by a human expert. Then, the embedding space is updated based on the human annotation. In the case the results are not related to the tongue ultrasound, the result will be annotated as an adverse sample in a different category. Fig. 4.10 depicts the process of updating the image label in embedding space using the human interactive segmentation.

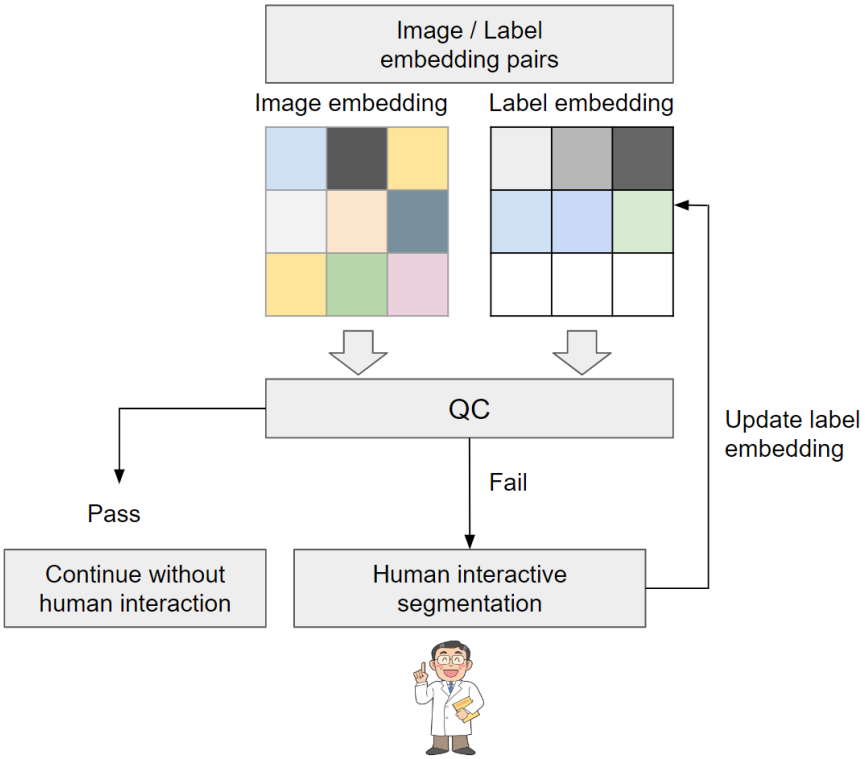


Figure 4.10: Visualization of the label embedding update for the user's interactive segmentation.

4.4 Vision Transformer Data Training

Fig. 4.11 depicts the data training paradigm for the Vision Transformer model. The data from ultrasound acquisition and transfer learning are split for 80% training and 20% validation. The K-fold validation, MSD, and Dice Score are mainly used to validate the training results (Please refer to Section 6.1 for evaluation measures details). There are two stages for the training. First, the model weights were initialized to be trained on data from transfer learning, which has 14M images. This is important to stabilize the ViT model convergence and learn the edges and fine features as the dataset is generalized and not for a specific task. Second, the trained model in the first stage will be fine-tuned (update model weights) on the task-specific tongue ultrasound image to extract the tongue contour segment.

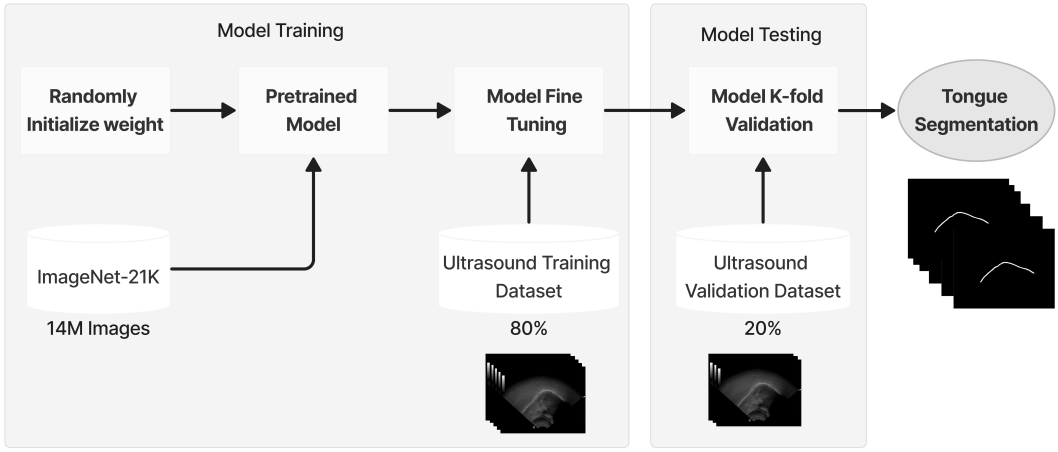


Figure 4.11: Data training paradigm using transfer learning and ultrasound data collected from the lab.

4.4.1 Data augmentation

In this research, we embraced the same augmentation strategy in [9]. The strategy utilizes standard augmentation using the following operations:

- Random cropping
- Image resizing
- Rotation

- Flipping
- Mirroring
- Random color distortion
- Random color blur

4.4.2 Data Ingestion from transfer learning

Vision-Transformer requires huge datasets to get accurate results compared to the well-known ResnNet-50. The Transfer learning dataset used in this research is trained on ImageNet-21k [223], which has about 14 Million images and 21 thousand classes. The benefit from the transfer learning model is that it can teach the algorithm about low-level features like image edges, textures, curvature and other image quality measures. This is important to have stable training and faster weight convergence. However, to improve the results, the model eventually fine-tuned on the target-specific task or the ultrasound images that are acquired in the lab.

Transformer performance is compared to the ResNet-50 when trained on ImageNet-21k [223], that have about 14 Million images and 21 thousand classes. On the other hand, the Transformer performance does not surpass the ResNet architecture when trained on ImageNet-1k [224, 225], which has 1.28 million image examples and one thousand classes. The best performance can be achieved when training the Transformer on the JFT dataset [222] that contains 300 Million images and 18 thousand classes. Unfortunately, JFT300 is not available for public use. Furthermore, it requires huge computation resources for data training. Refer to Figure 3.16 for comparative study details of different models.

4.4.3 Ablation study and results discussion for TongueTransUnet

The tongue contour segmentation results are assessed using the MSD and compared to the other techniques in different literature. Table. 4.1 shows the comparative analysis of average MSD and standard deviation for each result. However, the results are close to each other, but this is not the ultimate judgment. The reported numbers, as claimed in their original literature, and during the evaluation, we found that when those algorithms were tested on a wide range of datasets, the

algorithm failed as they were not generalized to a wide range of different features and ultrasound images. The proposed architecture succeeds in generalizing the model for any ingested data. Even in the worst-case scenario where the ultrasound image is noisy, and the tongue contour structure is heavily deformed human manual feedback is added to annotate the image, which boosts the model performance in the case of a low confidence score.

Table 4.1: Comparative study of MSD results of different methods for tongue segmentation. Results in mm (Each pixel= 0.295mm).

Method	$MSD \pm std$ **(mm)**
Research article	0.86 ± 0.11
EdgeTrack [13]	6.67 ± 3.93
Particle filter [15]	1.69 ± 1.1
CNN-based tool [117]	0.959 ± 0.58
TongueTrack [14]	3.48 ± 1.5
Autotrace [111]	2.61 ± 1.22
Biomechanical model [110]	0.71 ± 0.8
Automated robust contour tracking [105]	1.195 ± 0.286
Real-time automated tongue contour tracking [230]	0.91 ± 0.295

While generalizing the model is essential, it has challenges. The main issue is the computation complexity. The model performance is evaluated and assessed by proposing the desired region of interest of the only necessary image patches that may relate to the desired tongue contour area. The proposed methodology built a generalized ground truth data in embedding using a few hundred high-quality samples from human expert labels to serve as a reference to judge the quality of any newly ingested data.

Fig. 4.12 visualize the color scale of the average patch importance for each 64 patches from the ultrasound images. The figures show each patch in a gray-scale heat map that shows the importance of patches around the tongue contour dynamic movement area. The patch importance was calculated by taking the average cosine similarity score of each patch from the same area of extracted tongue contour at randomly selected 300 images. The patches on the image corners or edges have less effect. The results show that eliminating some image patches from image processing is possible without affecting the final results. This is important

to increase the model efficiency at scale while processing large volumes of ultrasound video at less computation and storage, which reduces the overall monetary cost.

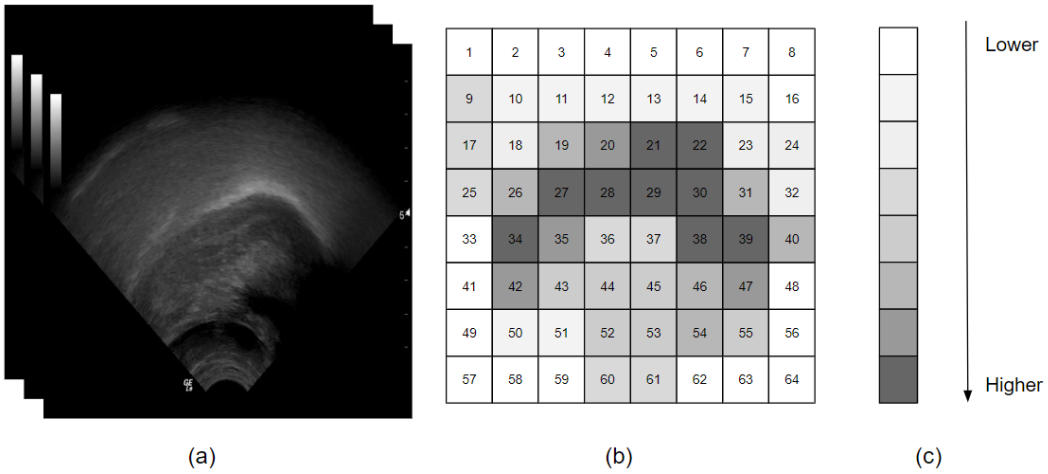


Figure 4.12: Visualization of patch importance of the average frames that affects the final image segmentation in ultrasound images. (a) The ultrasound image. (b) The color scale of the average patch importance of the 64 patches. (c) The color scale.

Table 4.2 shows the average MSD and Intersection over Union (IoU) compared to the percentage of total image patches used for training. The least important patches were removed at different percentages to minimize the computation complexity and focus only on highly important patches that relate to the desired region of interest. The average value computed by aggregating the similar patches values with similar index values and take the average to identify the feature importance. This help to classify the features and identify the dynamic pattern of the feature and contour displacement behavior.

The results show that the segmentation performance is almost the same as using 100% of patches as the same if we only use 85% of them. At the same time using 68% which is the mean of the highly important images are sufficient to provide good acceptable images to save more space.

The quality of the images degrades gradually while reducing the size of the utilized image patches. While the model hallucinates or generates unstable output if we use less than 25% of the image patches.

IoU and MSD results are different as the MSD resembles the normalized means

sum of distances between the ground truth and segmented contour lengths. At the same time, the IoU is derived by the intersection area of the segmented contour with the ground truth divided on the segmented contour and ground truth area union. While the proposed algorithm has many advantages, it also has a few limitations due to the nature of technical foundations.

Table 4.2: Ablation study of model performance concerning the percentage of patch used in image training.

Patch usage percentage	**Avg. MSD $\pm$ std (mm)**	**IoU (Jaccard)**
100	0.86 ± 0.11	0.94
85	0.81 ± 0.11	0.91
68	0.77 ± 0.11	0.88
50	0.56 ± 0.11	0.67
25	Model hallucinate	0.31

The TongueTransUNet architecture inspired by the TransUNet model. The novelty of this work as it focuses on managing the quality control of data flow through the entire life-cycle from model training until deployment. The enhancement over the basic implementation is to reduce the size of the required dataset by selecting high-quality and relevant information only.

The quality check is composed of three stages. In the first stage, an image similarity check using contrastive loss is used as a content filter for any image input. Second, Once the image passes the first stage the MSD segmentation accuracy and shape consistency analysis. The third stage is verifying the total segmenting accuracy score and if it does not pass the defined threshold the algorithm asks for human verification to do a manual check.

TongueTransUNet segmentation results have a small margin which makes it a stable and reliable architecture to segment the tongue contour during speech. The results also show that utilizing 85% of image patch compared to the original image size is the most efficient patch size that preserves image details and minimizes the computational complexity. TongueTransUNet processes the image in embedding space which gives more flexibility to extract features and analyze them in a multi-dimensional space. While TongueTransUNet has advantages at the same time it has some challenges. Below is a list of the main advantages and challenges for the TongueTransUNet.

Advantage of the TongueTransUNet:

- Do not need a huge dataset (in Millions), like traditional ViT.
- Avoid black-box view of the features in embedding space. The algorithm arranges each feature in its correct position in embedding space by using the contrastive loss and semi-fixed cluster location for features in the same category.
- Uses only necessary features with high attention scores and meets the quality control set of measures.
- Ensure that the segmented contour is usable and clinically valid using shape consistency verification.
- The QC relies on shape consistency and the normalized mean sum of distances accuracy measure that considers the variation of tongue contour lengths of different subjects to have normalized and valid comparative analysis.

Limitation and challenges of TongueTransUNet method:

- The core of this methodology relies on having a domain expert generate the reference results to shape the reference of ground truth embedding.
- The computation cost of the contrastive loss increases while the number of samples increases.
- The contrastive loss performance relies on the batch size. The larger the batch, the better, but the computation time increases.

Recommended future work for the TongueTransUNet:

- Using a multi-modality approach to process synchronous sound, text, and image data.
- Employ large language models to provide real-time sentiment analysis for the speech-text translation.
- Use predictive models to detect speech behavior and sound gestures.

Chapter 5

ZTCloudGuard: Zero Trust Context-Aware Access Management Framework to Avoid Medical Errors in the Era of Generative AI and Cloud-based Health Information Ecosystem

This chapter extend the previous work to harness the AI-based system as a feature extractor to implement a context aware system that verify the integrity, availability and confidentiality of the data journey within distributed telehealth system. The chapter presents a design of ZTCloudGuard which is a framework that uses a semantic and syntactic analysis of the users, devices and data output. These analysis utilized to implement a scoring model which can be used for managing the access in the complex healthcare information system.

5.1 Introduction

The access management context-aware system is designed to prevent or alleviate data breaches that are either from external intrusion or an internal error within the healthcare cloud information system. The focus of this research on alleviating

the medical errors, anomaly detection and wrong data entry.

Controlling access to data in healthcare systems is a major challenge for any service provider. While promoting the idea of smart hospitals and telehealth, it is required to look deeply into the existing regulation and access control systems to be more valid in the context of using new technologies like IoT devices, Cloud, AI, Blockchain, Quantum computing and 5G networks.

The Health Insurance Portability and Accountability Act (HIPAA) is patient information compliance in the United States, and The Personal Information Protection and Electronic Document Act (PIPEDA) is the Canadian version of regulating the healthcare information system. In Europe, the General Data Protection Regulation (GDPR) compliance regulates the sharing of information in healthcare. To adapt to the new advancement in technology, the dissertation proposes an access control system that can manage the patient information system within the complex structure of healthcare systems.

There are different types of smart or IoT devices in the healthcare environment.

- Patient monitoring devices (ECG, EEG, PPG, Blood Pressure, Temperature....).
- Telehealth Consulting (Remote doctor, chatbots, portable devices and handheld devices).
- Medical Imaging Systems.
- Robotics and Virtual Reality.
- Other smart devices at the hospital (Tv remote, voice assistance, washing machines).

5.2 Background and related work

Processing large volumes of data within the healthcare ecosystem makes medical report automation challenging and error-prone. To validate the input and output for the healthcare system data flow we have to ensure confidentiality, availability, and integrity. Implementing efficient access management is essential for data confidentiality. At the same time, an AI-based context-aware system can be used to validate data integrity [231, 232]. While a resilient system [233] is important to

make the system available when it is needed, a cloud-based system could be the recommended solution.

The existing healthcare information systems rely on the HL7FHIR standards [234] to define the communication and security access control system in healthcare. There are three main subsystems within the HL7FHIR:

1. **Authentication:** To verify the user.
2. **Access Control Engine:** To decide which FHIR controls are allowed for the user using the **CRUD** method (Create, Read, Update, Delete).
3. **Audit log:** to record the actions and any suspicious system intrusion.

At the organizational level, the access control system has three main common types within the health information system:

- **RBAC:** Role-based access control [235], [236].
- **ABAC:** Attribute-based access control [237, 238].
- **CML:** Modern cloud-based machine learning access control [239, 240, 241].

The **RBAC** and **ABAC** are the standard access control systems in healthcare. They are used widely in the traditional healthcare infrastructure setup for managing access control within the hospital perimeter. The **RBAC** manages the access based on the user role and grants permission based on the **CRUD** or **HTTP** method. **RBAC** is complex and considers different factors like users (operator, patient), roles, permissions, resources objects, and context of the data access [242, 234]. Please refer to Table A.2 in Appendix **??** for more information on the role-based access system factors. The role-based access management system has limitations that make it less effective in a complex modern healthcare environment. The **RBAC** is time-consuming and requires manual work to adjust rules and policies, which makes it less effective in real-time access management that has too many factors in a complex cloud-based environment.

The **ABAC** [243, 244] Based on predefined policies and conditions, **ABAC** grants system resources or objects access based on specific data attributes. For instance, compiling with regulations, the patient identification information can not be accessed without having the patient consent attribute to process the data. On the

other hand, the **ABAC** has a considerable amount of challenges. It is not efficient at big-data applications to limit its scalability. It is also time consuming and requires a considerable amount of resources to make it inefficient in a dynamic and modern globally distributed healthcare system.

In the traditional access control for healthcare information system, the user typically sends a request to the server through the REST API gateway. The server then sends a request for the REST API to verify the user information to grant the required **CRUD** operations based on the predefined rules and policies.

The modern cloud-based access control systems rely mainly on the zero trust principle that analyzes everything in the network and does not grant trust to any entity, either a user or device, for data access without passing a set of conditions that are defined by the organization policy. However, the main challenge is to identify what attributes or data context that will be considered in the policy without compromising the quality of the provided service. Mitigating the risk within the network is also essential to evaluate the access decision of the users [245].

Defining attributes and enabling zero-trust features requires utilizing advanced AI algorithms [246]. Among these algorithms, computer vision is essential for analyzing medical imaging [11, 247]. It has been used for different applications like ophthalmology [248]. Natural language processing (NLP) is useful for understanding the language details and predicting diseases [249]. NLP is also an effective tool for medical report processing [250]. NLP getting more attention in the area of privacy preserving and medical reports anatomizing [251].

Voice recognition is also part of digital health transformation and automation. Sound processing is vital for building context context-aware systems to validate data integrity. Speech can be used in anomaly detection applications and emotion recognition for people with special considerations and elderly people [252, 253].

Recently large foundational models that include huge datasets for image, text, and sound. Most of the current research is for building multi-modality systems. This type of model can help in processing complex unstructured data. Using large language models for healthcare queries was evaluated in [254]. Ingesting and analyzing electronic health records (EHR) are investigated in [255]. Several language models used for healthcare information systems from technology to ethical practices are surveyed in [256]. LLaMA and GPT-4 language models are the two common general-purpose language models. In the medical field accuracy is vital a specialized language model Med-PaLM [257] introduced by Google and trained on

high-quality medical data and meets expert level for answering medical questions. Large models trained on high-quality data take place in the healthcare industry as it alleviates the cost and minimizes technical and human-related errors.

Within the healthcare information system, there are different data sources used in the decision engine that act as a brain for the centralized healthcare information system. Medical images are typically stored in the Digital Imaging and Communications in Medicine (DICOM) format.

To store DICOM images the Picture Archiving Communication System (PACS) is used so the raw data can be further processed. Patient records are stored in Electronic Health Records (EHR). EHR communicates with other devices through High-Level Seven (HL7) and Fast Healthcare Interoperability Resources (FHIR) communication protocols. For other medical devices either portable, handheld, or hospital-based IoT devices the cloud-based infrastructure processes them in either real-time or batch processing mechanisms. Fig. 5.1 illustrates the main data sources that act as a backbone of a cloud-based healthcare information system that can be utilized for minimizing medical errors.

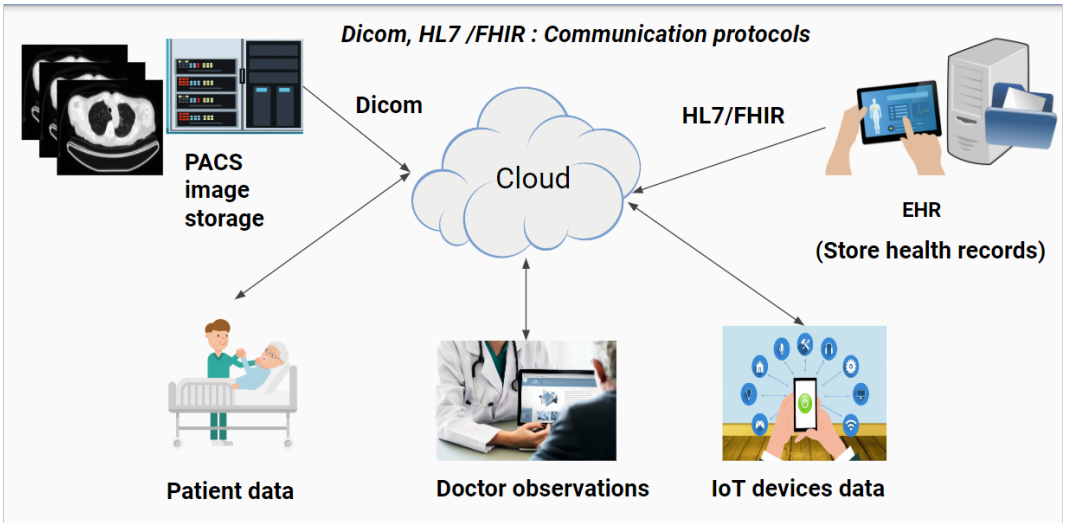


Figure 5.1: Visualization of the source of healthcare-related main information within the cloud-based system.

5.3 Method

In this section, the research methodology is explained in the following subsections. Sec. 5.3.1 provides an overview of the high-level architecture for the zero-trust context-aware system. Sec. 5.3.2 highlights the main pillars of the trust cycle for the zero-trust system. Sec. 5.3.3 evaluates the trust between different trust cycle attributes. Sec. 5.3.4 explains the hierarchical process of the decision engine.

5.3.1 Overview of the proposed zero trust framework for access management

The architecture design for the proposed context-aware access management frameworks is depicted in Fig. 5.2. The proposed access control system considers the zero-trust context-aware system to manage and analyze the data journey from the user of medical IoT devices endpoints to the cloud resources destination.

The proposed framework is classified into three main layers, as listed below:

- **Cloud input sources:** This layer is the front-end gateway for the main input source from users, devices metadata, and the context of data output either stored in the database or ingested in real-time streaming.
- **Cloud decision engine:** This is the centralized layer that acts as a brain for the decision engine. Build a chain of trust for each component based on the trust scores. There are two scores: critical trust (CT) and bond trust (BT). Then, the engine encodes the context attributes for further analysis at a hierarchical level. In the end, it grants the final access decision, operations, and constraints based on the analysis.
- **Cloud resources:** This is the back-end layer for the zero-trust ecosystem. The main components contain the cloud computing and storage resources that are used to process and store the metadata in the healthcare database.

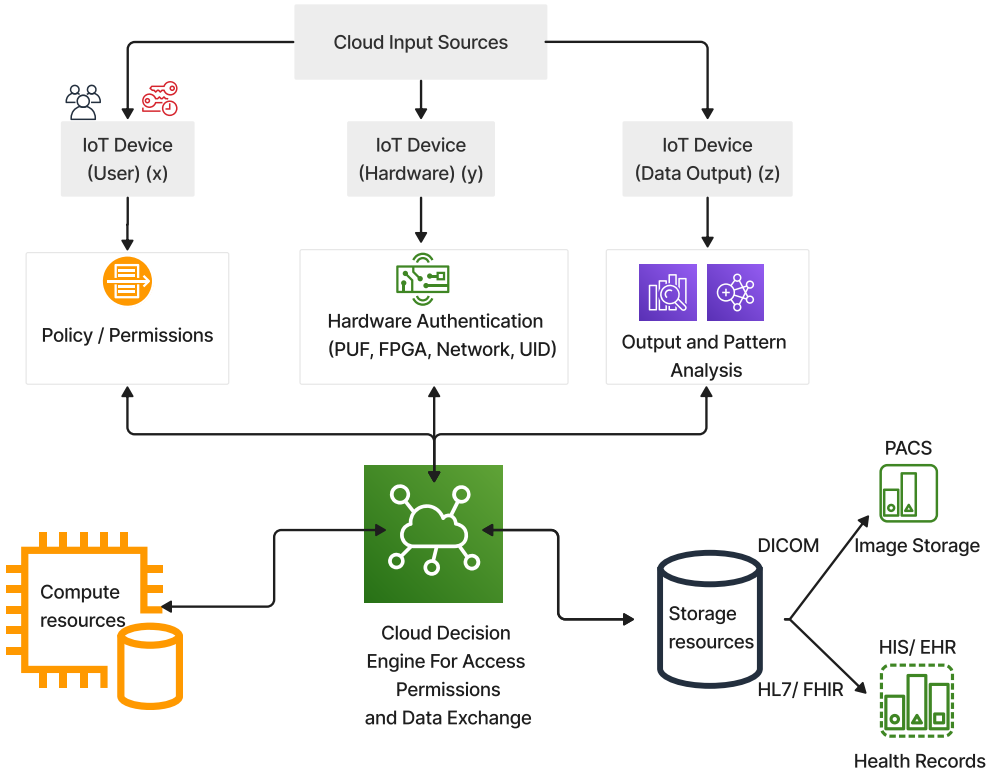


Figure 5.2: Representative of the proposed access control functional diagram within the healthcare cloud-AI ecosystem.

The following subsections explain the components of the proposed framework in detail.

5.3.2 Trust cycle pillars

The proposed system harnesses the zero trust context-aware system to manage access from the cloud input sources. The zero trust principle is based on utilizing all available data points for access management, including user identity, location, device health, services, workload, or data classification. There are three main components for the context-aware cycle that consider the context of the zero-trust five elements. Fig. 5.3 depicts the three components of the trust cycle. Who is the user, which device is used, and what is the output?

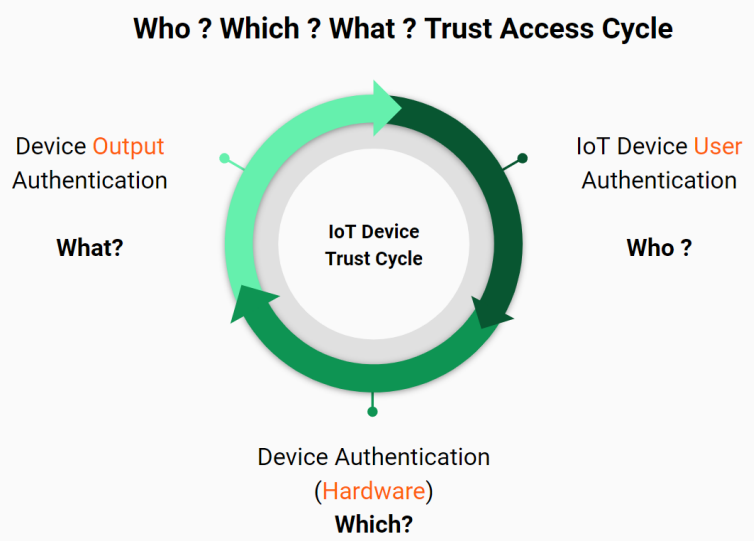


Figure 5.3: Trust Cycle of the proposed access control framework.

The trust cycle has five elements that are pillars of the proposed zero-trust principle.

1. User (Identity)
2. IoT device (Hardware)
3. Network (Device connection)
4. Application workload (Output patterns and scale)
5. Data (Output transaction context)

Where the identity relates to the user, the IoT device, and the network related to the hardware component. The application workload and transaction context relate to the output component.

5.3.3 Trust assessment

Building zero trust requires defining a set of attributes from different categories to verify the trust cycle. Zero-trust eco-system needs to be verified through a continuous trust cycle by implementing a series and chain of trust to assess semantic and syntactic relationships between the cloud input sources from users, devices,

and output data. The chain of trust is important to decide what level of access can be granted and to deny access if the connection is below the threshold of an acceptable trust score.

Fig. 5.4 illustrates the chain of trust and the assessment scoring criteria within the cloud ecosystem. The proposed framework constructs two assessment scoring criteria to manage the access of distributed medical devices. First is the critical trust (*CT*), which relies on cloud-native micro-services. Second is the bond trust (*BT*), which is a proposed scoring schema to manage access control, as explained below. *BT* uses machine learning pre-trained models to analyze the semantic and syntactic attributes from the trusted and authorized change of zero trust pillars 5.3.2, that are related to uses, devices, and data.

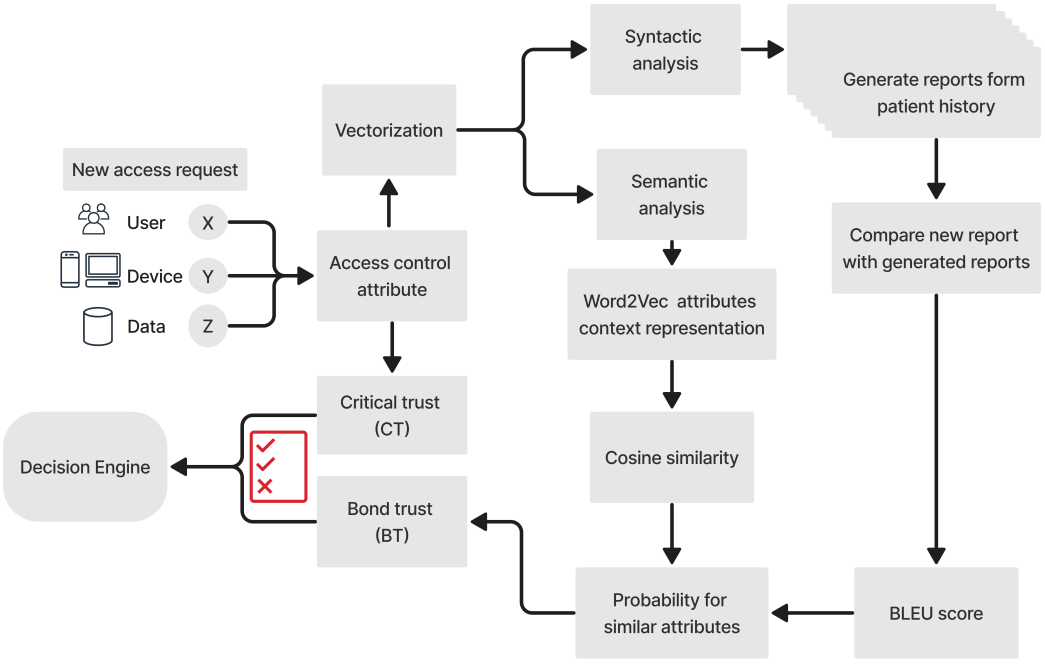


Figure 5.4: Illustration of the decision engin in the proposed framework of a continuous chain of trust based on the accumulated trust.

Critical Trust (*CT*): *CT* is the initial evaluation and scoring criteria to grant access to the cloud ecosystem. This assessment grant is preliminary and not for direct connection to the back-end resources for storage and computation. *CT* is important as it acts as an additional layer of security to separate user access control from the actual dataset resources. *CT* is evaluated using cloud-based micro-services. There are four main attributes for the critical trust score. Cloud-based micro-services like

authorization, authentication, logging, and encryption are digitized to derive the final CT score as per Eq. (5.1).

Each micro-service attribute is given a logical value, 1 or 0. Then, these micro-services logical values are multiplied by a scoring factor (S_i) based on its importance that can be set by the system admin. The cloud decision engine grants access status to **allow** for trusted authority, **verify** if more information is needed, and **deny** for non-trusted access requests.

$$CT = S_1 \times A_1 + S_2 \times A_2 + S_3 \times A_3 + S_4 \times A_4 \tag{5.1}$$

where A_1 is authentication and its scoring factor is S_1, A_2 is the authorization and its scoring factor is S_2, A_3 is the encryption and its scoring factor is S_3, A_4 is the logging and its scoring factor is S_4. Table 5.1 provides an example of critical trust score evaluation using different scoring factors and micro-services logical values.

Table 5.1: Examples of critical trust score assessment.

ID	A_1	S_1	A_2	S_2	A_3	S_3	A_4	S_4	**Critical trust score**	**Access status**
D1	1	0.3	1	0.4	1	0.2	1	0.1	0.9999	Allow
D2	1	0.3	0	0.4	1	0.2	1	0.1	0.6	Verify
D3	0	0.3	0	0.4	0	0.2	0	0.1	0	Deny

Bond Trust (BT): Once the transaction is passed the critical trust assessment, the bond trust will evaluate the relationship to other resources to build a trust cycle for only authorized and highly trusted actors to make sure that the data or resources are accessed by designated people based on the organization policy or rules. Calculating bond trust is more complex and depends on different aspects. BT has two main assessment criteria. First is BT_A, which assesses the semantic relationship between each individual attribute stored in the health care information system. Second is BT_B, which assesses the syntactic relationship between the set of candidates in a generated health report. The reason for using these two measures as the first one is that it is essential that each attribute has meaningful meaning and is related to similar attributes compared to the pre-trained ones. The second one is essential to guarantee that the attributes in the generated report are in the context of the patient's history to ensure that the report is highly likely related to the same patient and, hence, is not diagnosed falsely with the wrong case.

The proposed assessment of BT_A uses Attribute2Vec representation that is based on the pre-trained word2vec model [139], [258]. The Attribute2Vec maps the attributes and their synonyms words that have the same context from the user (x), hardware (y), and output (z) attributes stored in their electronic health records. The skip-gram methodology [259] is used to derive the attributes with the same context, and we suggest in this framework using the first three words with the highest context probability. The advantage of using this assessment technique is to generalize the model by accepting a wide variety of attribute descriptions in a global context. word2vec is valid for different languages and dialects. For example, it has been used by Altibbi.com [260] to train 1.5 Million medical consultation questions in the Arabic language. We recommend using a matching engine on the Vertex AI platform at Google Cloud to make sure the word embedding and vector similarity matching process is efficient and reliable.

Fig. 5.5 depicts the process of assessing bond trust. The input has the three attributes of users, devices, and output. The hidden layer extracts features and the $SoftMax$ is used for predicting probability to extract a set of similar attributes that have the highest probability. In this research, we selected the highest three attributes. Eventually, the cosine similarity is used to predict the relationship between attributes from different categories (x, y, and z). The bond trust scoring is used to derive the final score to decide accepting or reject the attributes based on the pre-defined threshold.

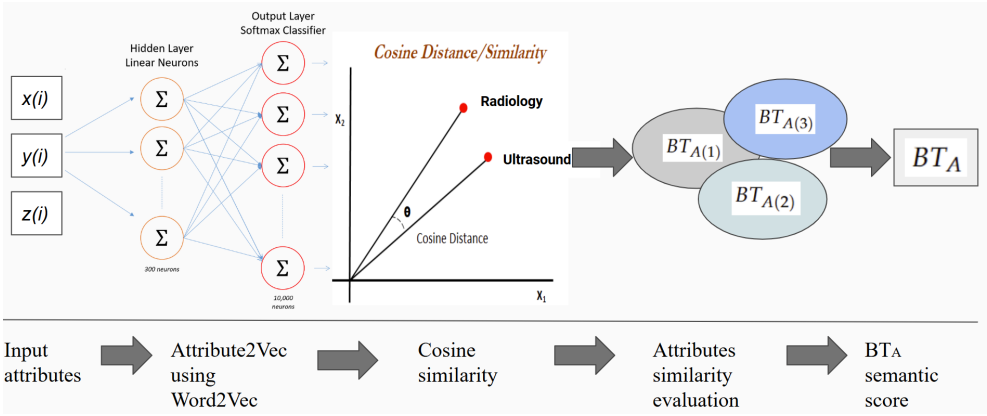


Figure 5.5: Semantic trust assessment using attribute2vec based on word2vec model. Where $BT_{A(1)}$, $BT_{A(2)}$, and $BT_{A(3)}$ are the set of bond trust between the three input sources x,y, and z. BT is the final bond trust score.

Cosine distance is used in Eq. (5.2) to predict the similarity probability of the context of attributes of x, y, and z.

$$\textit{Similarity (A,B)} = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\left\|\vec{A}\right\| \left\|\vec{B}\right\|} \tag{5.2}$$

where $\vec{A} \cdot \vec{B} = \sum_{i=1}^{N}(A_i \times B_i)$ is the dot product between two vector attributes $\vec{A}$ and $\vec{B}$. At the same time, $\left\|\vec{A}\right\| = \sqrt{\sum_{i=1}^{N}(A_i)^2}$, $\left\|\vec{B}\right\| = \sqrt{\sum_{i=1}^{N}(B_i)^2}$ are the L2-norms of the attributes $\vec{A}$, $\vec{B}$ respectively. While θ is the angle between the two vectors.

The highest probability attribute vectors between x, y, and z will be used to derive the bond or semantic mutual relationship in three bond-trust scores sets $BT_{A(1)}$, $BT_{A(2)}$ and $BT_{A(3)}$ for the relationship between xy, xz and yz. Where $BT_{A(i)}$, the bond trust sets are derived using two inputs:

A. Cosine similarity logical evaluation: Algorithm 2 is used to assign a logical value for the cosine similarity between two attributes. There is a given value of either one or zero based on the relationship of the attributes x,y, and z. The value is assigned based on the threshold of the angle θ between the two attributes. Eq. (5.2) is used to derive θ using cosine similarity between the attribute vector product for the given (i) index or position for similar context attributes. The algorithm produces a set of three logical values $Sim_A(\vec{x}_i, \vec{y}_i)$, $Sim_B(\vec{x}_i, \vec{z}_i)$, $Sim_C(\vec{y}_i, \vec{z}_i)$, for each given index($i$).

B. Weight: The weight is calculated using the GloVe model [261] of word embedding to consider the co-occurrence of the attributes in a global representation context of the healthcare database. The weight is based on the conditional probability of attribute occurrence or importance as in Eq. (5.3).

$$w_i = \frac{P_{BA}}{P_B} \tag{5.3}$$

where w_i is the probability of word B occurrence in the context of the word A of a given (i) index of two semantic or syntactically similar attributes.

The three scalar values of $BT_{A(1)}$, $BT_{A(2)}$ and $BT_{A(3)}$ are stored in BT_A as shown in Eq. (5.4). Where BT_A is a 1×3 vector.

$$BT_A = [BT_{A(1)}, BT_{A(2)}, BT_{A(3)}] \tag{5.4}$$

Algorithm 2 An algorithm for proposed cosine similarity logical evaluation process

Input: User (x), Device (y) , Output data (z), Angle threshold (Th_θ)

if $\theta_{xy} \geq Th_\theta$ **then**
    $Sim_A(\vec{x}_i, \vec{y}_i) = 1$
else if $\theta_{xy} < Th_\theta$ **then**
    $Sim_A(\vec{x}_i, \vec{y}_i) = 0$
end if
if $\theta_{xz} \geq Th_\theta$ **then**
    $Sim_B(\vec{x}_i, \vec{z}_i) = 1$
else if $\theta_{xy} < Th_\theta$ **then**
    $Sim_B(\vec{x}_i, \vec{z}_i) = 0$
end if
if $\theta_{yz} \geq Th_\theta$ **then**
    $Sim_C(\vec{y}_i, \vec{z}_i) = 1$
else if $\theta_{yz} < Th_\theta$ **then**
    $Sim_C(\vec{y}_i, \vec{z}_i) = 0$
end if
Output: $Sim_A(\vec{x}_i, \vec{y}_i)$, $Sim_B(\vec{x}_i, \vec{z}_i)$, $Sim_C(\vec{y}_i, \vec{z}_i)$

where BT_1 is the relationship score between the user (x) and hardware (y) and is derived using Eq. (5.5). BT_2 is the relationship score between the user (x) and output (z) and is derived using Eq. (5.6). BT_3 is the relationship score between the output (z) and hardware (y) and is derived using Eq. (5.7).

$$BT_{A(1)} = \sum_{i=1}^{N} (w_i)_{xy} \cdot Sim_A(\vec{x}_i, \vec{y}_i) \tag{5.5}$$

$$BT_{A(2)} = \sum_{i=1}^{N} (w_i)_{xz} \cdot Sim_B(\vec{x}_i, \vec{z}_i) \tag{5.6}$$

$$BT_{A(3)} = \sum_{i=1}^{N} (w_i)_{yz} \cdot Sim_C(\vec{y}_i, \vec{z}_i) \tag{5.7}$$

where w_i is a scalar weight that is used to scale the bond score for each attribute based on the importance of the feature at given (i) and derived by Eq. (5.3). N is the sequence number of attributes that are numbered based on the probability of their context relationship. Only each similar class attribute of user, devices, and output is multiplied by each other, and if they belong to the same category, the algorithm gives them either a 0 or 1 similarity score and then multiplies them by the scalar weight for that attribute. The step is then repeated for all attributes. The final multiplication is aggregated to have a final scalar number that resembles the combined similarity score for $BT_{A(i)}$.

$BT_{A(i)}$ vector is normalized in Eq. (5.8) using the $SoftMax$ function. The normalization process produces a new vector BTN of a 1×3 dimension.

$$BTN_i = SoftMax(BT_{A(i)}) = \frac{\exp(BT_{A(i)})}{\sum_j \exp(BT_{A(j)})} \tag{5.8}$$

The result is stored in Eq. (5.9) and has three scalar values that are between zero and one.

$$BTN_i = [BTN_1, BTN_2, BTN_3] \tag{5.9}$$

The first part of the bond score is calculated in Eq. (5.10) by aggregating the three normalized scores BTN_1, BTN_2, and BTN_3.

$$BT_A = BTN_1 + BTN_2 + BTN_3 \tag{5.10}$$

where BT_A is between zero and one. Zero is for non-matched attributes, and one

is for the highest attribute similarity match. Any number between zero and one requires an additional trust verification and re-assessment.

At the same time, BT_B is used to assess the similarity in the generated report text by evaluating the syntactic performance of the candidate report generated from the stored data in the healthcare information system. Unlike semantic analysis, the syntactic analysis is effective for evaluating a full report not just single word meaning. On the other hand, semantic analysis provides a wide context analysis using various probabilistic-related attributes. BT_B is inspired by the $BLEU$ score [262], which was originally designed by IBM for machine translation scoring evaluation as shown in Eq. (5.11).

$$BT_B = \min(1, exp(1 - \frac{reference - length}{output - length}))(\Pi_{i=1}^{n} precision_i)^{1/n} \tag{5.11}$$

where BT_B has two parts. First is the brevity penalty that compensates for the length of a short generated report. Second, the precision for n-gram candidates. The n refers to the number of candidates used to evaluate the score. The notation n typically is 4 and can be increased to include more restrictions for identifying medical errors. In the case of $n = 4$, the $BLEU$ score needs the candidate report to match the reference template by at least four attributes.

In the case of no patient history, the BT_B score is zero and less effective for the syntactic analysis. This scoring evaluation is meaningful when the patient has a previous history in the EHR. The final bond trust normalizes the summation of BT_A and BT_B to keep the value between zero and 1 ad in Eq. (5.12).

$$BT = \frac{BT_A + BT_B}{2} \tag{5.12}$$

5.3.4 Decision engine encoding and hierarchy

The decision engine for similarity scoring is built through encoding and a hierarchical process. There are two main stages for the hierarchical process. These stages are vital to make sure that the final decision follows logical flow based on a set of constraints. While the encoding Fig. 5.6, visualizes the two stages of encoding and hierarchy process for access control management.

Stage one: The decision engine performs the initial critical check for the end

point device or user request access from the server. The critical check is essential to guarantee that the endpoint components have passed the regulatory compliance and critical trust score threshold, Table. 5.1. An example of one of the main regulatory compliance that needs to be considered is the HIPAA, which lists 18 patient information identifiers [263] that are restricted from being shared without having consent from patients and meet all security guidelines within the healthcare information system.

Stage two: The decision engine encodes the attributes from devices, output, users, and critical trust in a 32-digit hexadecimal array. The 32-digit array is then analyzed to make the final decision. The final decision is also encoded to resemble the access level, operations, access resources, and constraints, see Sec. 5.3.5.

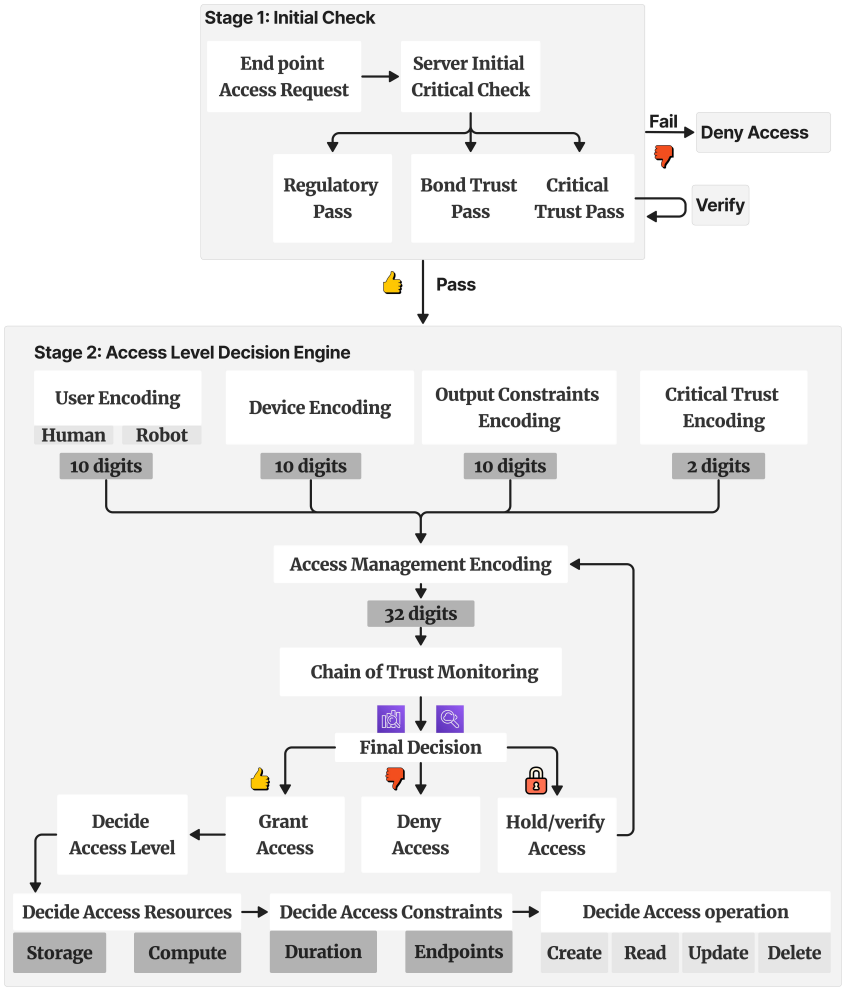


Figure 5.6: Representative of the access control engine decision hierarchy and the encoding.

Fig. 5.7 illustrates an example of the encoding criteria for the proposed three zero-trust components, user, device, and output, based on different attributes. Each component has a 10-digit hexadecimal value and a 2-digit value for each one of the five attributes. The importance of these attributes is to ensure that the access request belongs to the designated group, has a pre-defined access level and type, and passes the bond trust.

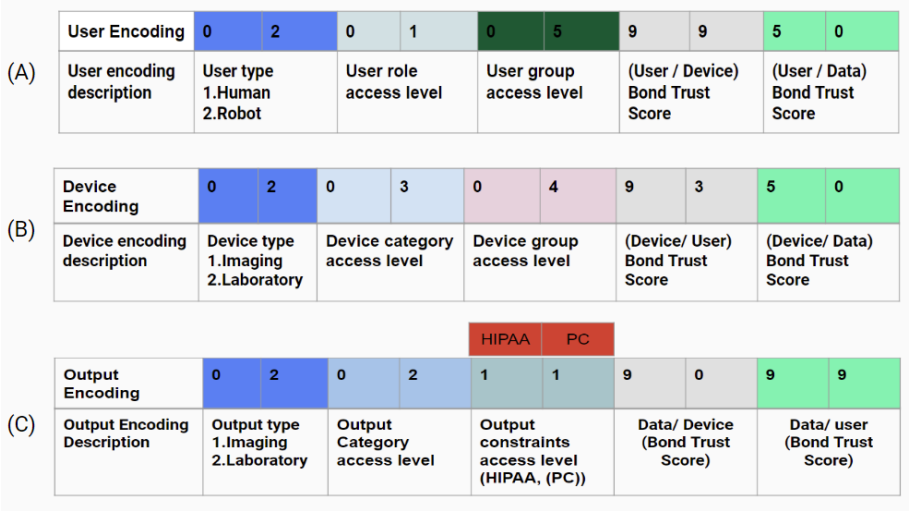


Figure 5.7: Proposed access control encoding. (A) User Encoding. (B) Device Encoding. (C) Output encoding. PC is the patient consent.

5.3.5 Final decision and access operations

The final decision has four main criteria and information, as listed below. Table. 5.2 depicts the final decision encoding information using hexadecimal digits.

- **Access level:** Decide the access level for each transaction.
- **Access resources:** Decide which storage and computation resources will be used.
- **Access constraints :** Decide what the access constraints are, such as duration, location, number of access trials, and size of data transferred.
- **Access operations:** Grant access based on CRUD or HTTP method. Operations also can be customized based on the organization preferences.

Table 5.2: Final decision encoding.

F	Decision	Hex	Encoding Description
F1	Access Level	2 digits	To specify the five access levels. Ex. 10 for access level L0.
F2	Access resources	6 digits	The first three digits are for compute resources, and the rest are for storage resources metadata.
F3	Access constraints	16 digits	8 digits for time, and the other eight digits for other constraints. Ex. 6421EC5F for 2023Y, 03M, 27D, 21h, 19mm, 59ss.
F4	Access operations	1 digit	For example, F is in hexadecimal to represent organization access of all operations

Algorithm 3 shows the logical process of the proposed framework. The framework has three inputs x, y, and z. The initial step requires passing the threshold for CT and BT that is specified by the system admin. Typically $CT \geq 99.99\%$, $BT \geq 0.7$, where each attribute in $BT_i \geq \theta$. If the score of CT and BT was zero, the access was denied, and any value between zero and threshold, the access should be verified again within a given time interval. The final access decision will be granted based on the assessment of the trust scores.

Algorithm 3 Proposed access management framework logical process

Input: User (x), IoT hardware (y) , IoT output data (z)
Trust assessment: Critical trust (CT), Bond trust (BT), Trust threshold (Th)

if $CT = 0$, $BT = 0$ **then**
    Deny access
end if
while $Th \neq 0$ **do**
    if $CT \geq Th$, $BT \geq Th$ **then**
        Initially accepts access
    else if $CT < Th$, $BT < Th$ **then**
        Verify access again
    end if
Output: Grant final access decision
end while

5.4 Experiment and results discussion

5.4.1 Dataset information

The dataset used in this experiment contains a synthetically generated set of attributes for users, medical IoT devices, and data. The generated data is synthesized using Synthea [264]. The synthetic dataset is used to fine-tune the word2vec model to enhance the result within a healthcare information system. Fig. 5.8 shows some examples of the generated dataset attributes of users, devices, and data that are used to prevent medical errors.

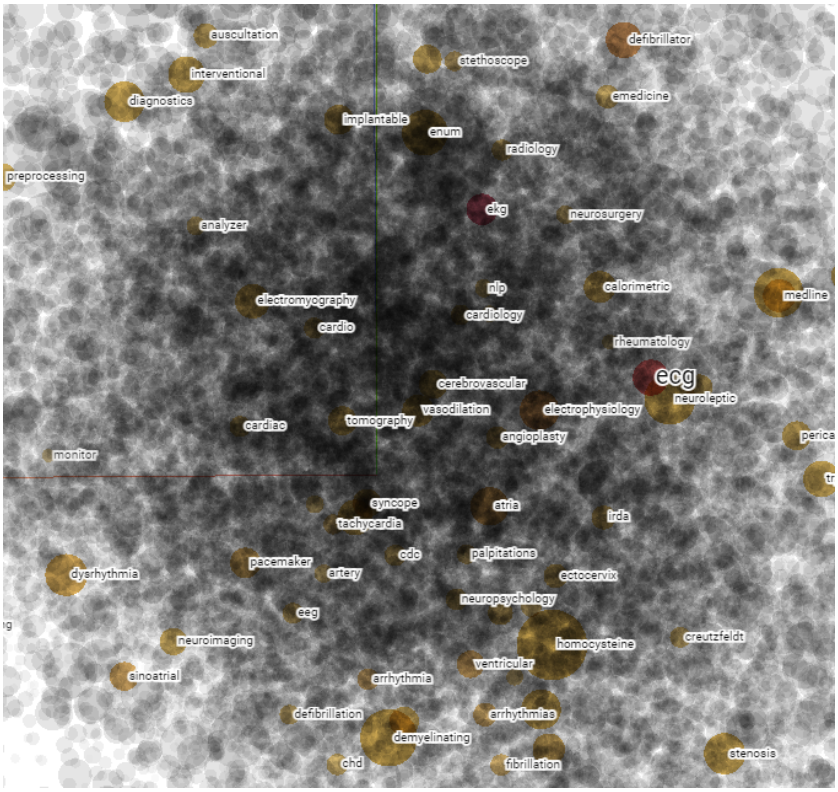


Figure 5.8: Example of selected attributes from the generated data using Synthea and fine-tuned word2vec pre-trained model.

The synthetic dataset is not limited to only a small subset of attributes. These attributes are extracted from large language models, and they include all attributes that are related to the healthcare information system for the main three categories user, device, or data. Examples of some of these categories are listed in Table. A.1.

In order to evaluate the syntactic information, we generated a sample of different patient reports based on a predefined template from the same set of attributes mentioned before. The generated templates generate all possible syntactic and semantic similarity reports that may be related to the patient based on the medical history. Fig. 5.9 depicts an example of a template used to generate the final report that includes information about users, devices, and data. There are different templates used according to the case and required information.

Template of generated prompt of medical record

A **patient name** with **patient Id** and **Insurance number** born on **birthday.** Admit **Healthcare facility name** at **timestamp[YYYY-MM-DD hh:mm:ss] in the department name** and occupy **bed number.** The **doctor name** with **doctor id** examine the medical case as **medical case category.** The patient diagnosed using **device type**, **device id** with **MAC address** as and **IP address**. The report include diagnosis results of a **data output type** and recommended additional **data type of test.**

Example of generated report using the template

A **John Doe** with **GC12346** and **IBN46792** born on **12/03/1995.** Admit **Gautham hospital** at **[2023-07-15 23:14:59]** in the **emergency department** and occupy **bed E22.** The **doctor Jane Doe** with **ID: 7987731** examine the medical case as **COVID-19.** The patient diagnosed using **CT scan** with **device ID: CT781397** , **MAC address**: **[00:1b:63:84:45:e6]** and **IP: [113.161.1.163]**. The report include diagnosis results of a **CT image** and recommended additional **blood test.**

Figure 5.9: Arbitrary example of generated text prompt from patient history record. The mentioned names are arbitrary examples and do not refer to any true identity.

5.4.2 Ablation study results and discussion

The ablation study is examined to evaluate the accuracy of identifying medical errors using the proposed model by examining the relationship between different attributes based on the critical and bond trust scoring. The study was conducted using 17625 attributes for users, medical IoT devices, and data output categories. The study shows that the F1-score is 93.5%, which means that the proposed methodology is valid for identifying the relationship between different attributes within the healthcare information system and alleviating any medical errors that may produce false medical reports. Fig. 5.10 depicts the confusion matrix of the experiment result.

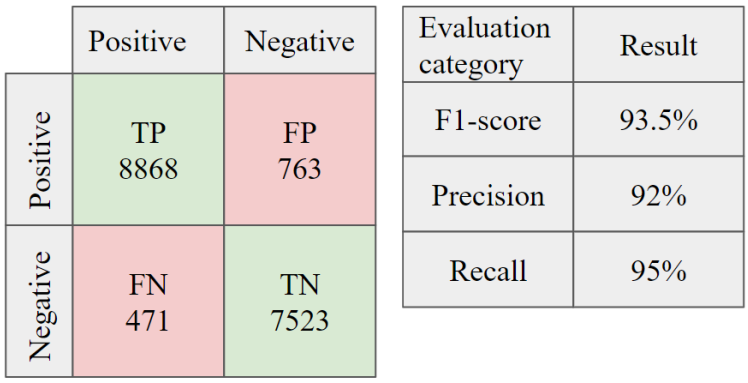


Figure 5.10: Confusion matrix for the ablation study on the accuracy of detecting medical errors by identifying the relationship between selected attributes. TP is true positive; FP is false positive; FN is false negative; TN is true negative.

Evaluating the final results has different criteria for semantic and syntactic information. Semantic information is used to evaluate the relationship of each word in the context of medical-related data. At the same time, syntactic information analysis is used to evaluate the corpus context of the newly generated report and compare it with the stored data in a healthcare information system. Table 5.3 lists some examples of the evaluation of the BT_A to extract semantic relationships from the healthcare information system for different medical specialty classes.

Table 5.3: The precision, recall, and f1-score for evaluating BT_A on a selected variety of speciality cases.

Speciality Class	Precision	Recall	F1-score
Radiology	0.880	0.880	0.880
Gynecology	0.773	0.840	0.805
Oncology	0.793	0.772	0.782
Dermatology	0.712	0.740	0.726
Cardiology	0.833	0.871	0.851
Urology	0.765	0.724	0.744
Emergency	0.865	0.834	0.849
Dentistry	0.79	0.77	0.779
Psychology	0.766	0.784	0.774

Table. 5.4 shows the effect of syntactic analysis using different measures on the final decision. As shown in the table, the 1-gram measure of the accuracy of detecting one word compared to the context of reference length from the stored data

in the healthcare system. While it has high precision, it is not accurate for making decisions as it does not account for the relationship with other attributes. The decision confidence increases gradually with reference to the n-gram rank as it has a more meaningful meaning.
The BLEU score is good for judging a corpus of attributes but performs badly on a single entry. In the case of syntactic analysis, it is more efficient to score the generated reports. However, it is not efficient for judging semantic information or detecting grammatical error sentences.

The proposed method got the best results as it accounts for semantic and syntactic information. The decision engine in the cloud generates different reports from the stored data that account for different synonyms, words, or attributes that are related to the stored data. It also can fix any grammatical errors in the entry and suggest an attribute within the same context. This gives the method a generalized capability to assess any new report or data entry within the healthcare information system through distributed users, devices, and sorted data. At the same time, the proposed method accounts for the security measure that requires authentication, authorization, encryption, and logging. Table 5.4 compares the confidence score of the proposed method with other scoring metrics that are used for syntactic analysis.

Table 5.4: Comparison of proposed scoring method and other metrics of syntactic analysis.

Metric	Decision confidence score
1-gram	26%
2-gram	33%
3-gram	47%
4-gram	66%
BLEU	71%
Proposed method	89%

The proposed framework focuses on the cloud-AI access control system by managing the access to the cloud resources for users, devices, and data. This is done by implementing a zero-trust context-aware system that analyzes the data for each transaction and always for the uses of the users, devices, network, workload, and data. The framework considers the information security model that

has three pillars of data confidentiality availability. Regulatory compliance is also part of context-aware access control. HIPAA is the most important compliance that identifies protected health information data. The protected data can not be used without following a series of security and privacy protection guidelines like patient consent, disclosure agreement, de-identifications, data encryption, and a well-managed access control system.

The framework also takes advantage of cloud-native micro-services to implement critical trust assessment criteria. To build a chain of trust between different attributes for each component, the framework proposes bond trust evaluation that is inspired by the large language models.

Table. 5.5 shows a sample of results for access management decisions based on the evaluation of the critical trust (CT) and bond trust (BT) assessments.

Table 5.5: Example of access management decision based on scoring evaluation.

$N_{Samples}$	CT (avg.)	BT (avg.)	Decision
402	0	0	Decline
267	0.99	0	Decline
224	0.99	0.5	Verify
259	0.99	0.9	Accept
110	0.99	0.83	Accept
479	0.99	0.79	Accept

While the zero trust context-aware system is robust against different situations, there are different challenges and limitations to implementing it in the healthcare industry, as follows:

- **Data Privacy and Security:** ML acts as a backbone of the zero trust access control system, which requires being trained on a considerably large dataset, the size required be in millions or even billions of parameters to get efficient results. Obtaining sensitive and accurate data is challenging due to privacy concerns for health information regulatory compliance, which may limit the accuracy of the system.

- **Complexity:** The complex healthcare IT infrastructure makes it difficult to implement and manage a zero-trust context-aware access control system. These systems need to be able to integrate with existing systems and appli-

cations, and they need to be able to handle the large volume of data that is generated in healthcare settings.

- **Cost:** Implementing a zero trust access control system requires an enormous investment in back-end infrastructure. The costs of implementing and maintaining these systems need to be balanced against the potential benefits, such as improved data security and reduced risk of data breaches, compared to the cost of investment.

- **Skills:** Zero trust principles rely on too many factors. These factors should be aligned with the current and most advanced technology, which requires highly skilled professionals who are always in demand due to the shortage of these skills in most employees.

Future considerations: The current research implemented a zero-trust context-aware system to minimize medical errors by analyzing the data context from users, devices, and data. The current model utilizes a fine-tuned Word2Vec model to analyze attributes. To improve the current algorithm, the recommended future work should consider more secure protocols like data encryption, blockchain technology, and using larger language models to improve model performance.

Using larger models like GPT4, Gemini, Mistral, Llama, and Claude can improve the accuracy of the current implementation. In addition, utilizing these models will increase the generalization capabilities as these models are trained on larger data sources. The recommended future work for securing medical IoT devices by employing the Physical Unclonable Function (PUF) to authenticate the hardware for the telehealth distributed devices [265]. PUF is getting more legitimate attention and adopted by USA presidency administration as a recommend technology to secure IoT devices.

5.4.3 Example of device output context-aware system

Fig. 5.11 presents an example of using the proposed protocol to analyze the ultrasound system output using the zero trust context-aware system to manage the access control of the device data to the cloud server.

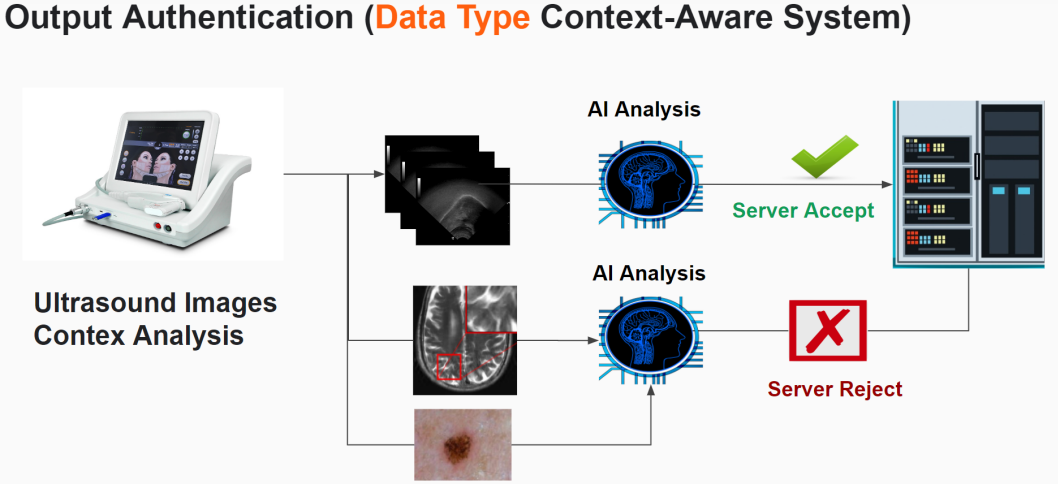


Figure 5.11: Example of the context-aware system of ultrasound device output analysis.

Chapter 6

Comparative Analysis of Tongue Segmentation and Qualitative and Quantitative Evaluation Metrics

This chapter presents the main evaluation metrics that are used to evaluate tongue contour tracking. The chapter also is an extension of the comprehensive review in Chapter 2. The main contribution of this chapter is that it proposes a unique quantitative and qualitative evaluation of tongue-tracking data integrity and does not rely only on quantitative evaluation like other methodologies.

6.1 Evaluation Measures for Tongue Contour Extraction Using Ultrasound

Different techniques are used to evaluate the accuracy of the extracted tongue contour. These techniques use manual or fully automatic extracted tongue contours as reference data. The typical and most accurate methodology to compare the result is by measuring the difference between the segmented tongue contour in the proposed methodology with the extracted ground truth contour. The ground truth data are labelled manually by a human who is specialized in using ultrasound systems. Some researchers use automatically extracted data to validate their results. However, automatically extracted data are less accurate than manual ground truth data. However, they are used when dealing with a massive dataset, as it is time-consuming to produce manual data. Whether the reference data are extracted

manually or automatically, the methodology to measure the difference between the extracted and the referenced data is similar and specific measures indicate the accuracy of the methodology. Some measures are valid for either traditional or machine learning techniques, and some other measures are only valid for machine learning techniques.

6.1.1 Mean Sum of Distances (MSD)

The mean sum of distances measure is adopted widely as an evaluation measure for tongue tracking and segmentation; it was proposed by [20]. The mean sum of distances is derived by comparing the automatically extracted tongue contours by the algorithm to the ground-truth-extracted contours by measuring the distances in two main steps. First, the minimum distance between each element on the algorithm-extracted contour and the nearest element on the ground truth is determined. Second, from the ground truth contour, the minimum distance for every point is measured against the nearest point on the algorithm-extracted contour. The sum of the minimum distances from these two steps is divided by the total number of elements in the ground truth and automatically extracted contours to normalize the results. Equation (6.1) shows the formula for the MSD.

$$MSD(U,V) = \frac{1}{m+n}(\sum_{i=1}^{n} min_j(\mid v_j - u_i \mid) + \sum_{j=1}^{m} min_i(\mid u_i - v_j \mid)) \qquad (6.1)$$

where (n) is the contour length of the ground truth, and (m) is the length of the automatically extracted contour, while (v_j) is the manually extracted contour (ground truth) data points, and (u_i) is the automatically extracted contour datasets. On the other hand, (min_i) and (min_j) illustrate the nearest distances between each point on the contour and the nearest point on the other contour, respectively. The MSD has a significant advantage because the length of two contours is not comparable, and other comparison methods such as the mean sum of errors and norm are inappropriate. The MSD is measured in pixels and then converted to millimetres by assuming that each pixel is 0.295 mm [15, 8].

6.1.2 Shape-Based Evaluation

Tongue contour image segmentation techniques are evaluated by the shape-based triangle measure proposed by [266] . Equation (6.2) is used to measure the

curvature, while Equation (6.3) describes the asymmetry of the tongue contour.

$$K = \frac{||CD||}{||AB||} \tag{6.2}$$

$$V = \frac{||AD||}{||DB||} \tag{6.3}$$

This evaluation measure considers the asymmetry and curvature of the tongue shape. $||CD||$, $||AB||$, $||AD||$, and $||DB||$ depict the segment lengths that are shown in Figure 6.1.

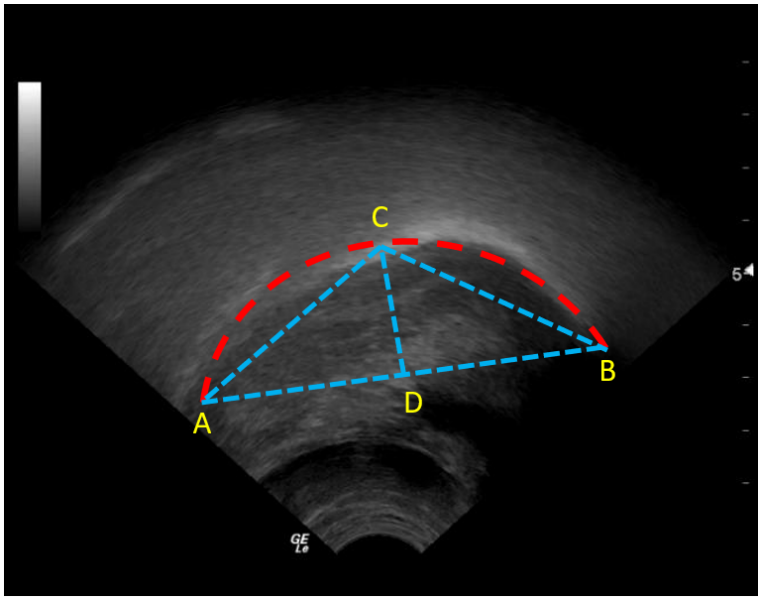


Figure 6.1: Shape-based evaluation measure. Point (**A**) is on the dorsal tongue part, point (**B**) is the point on the tongue tip, point (**C**) is the apex. Point (**D**) is the projection of point (**C**) on the (**AB**) line. [8].

6.1.3 *K*-Fold Cross-Validation

Figure 6.2 shows the data validation on different folds or segments to maximize the model performance. The *K*-fold cross-validation method can be used to evaluate machine learning models' performance by comparing the training and validation datasets [267]. The *K*-fold process can be done by partitioning the complete datasets into a number *K* of segments. For instance, the typical practice of model validation uses 80% of the segments for data training and 20% for validating the

data. The *K*-fold cross-validation shuffles between the *K* segments to reassign different subsets into the validation and training segments. The final performance is evaluated by computing the mean sum of the *K*-folds.

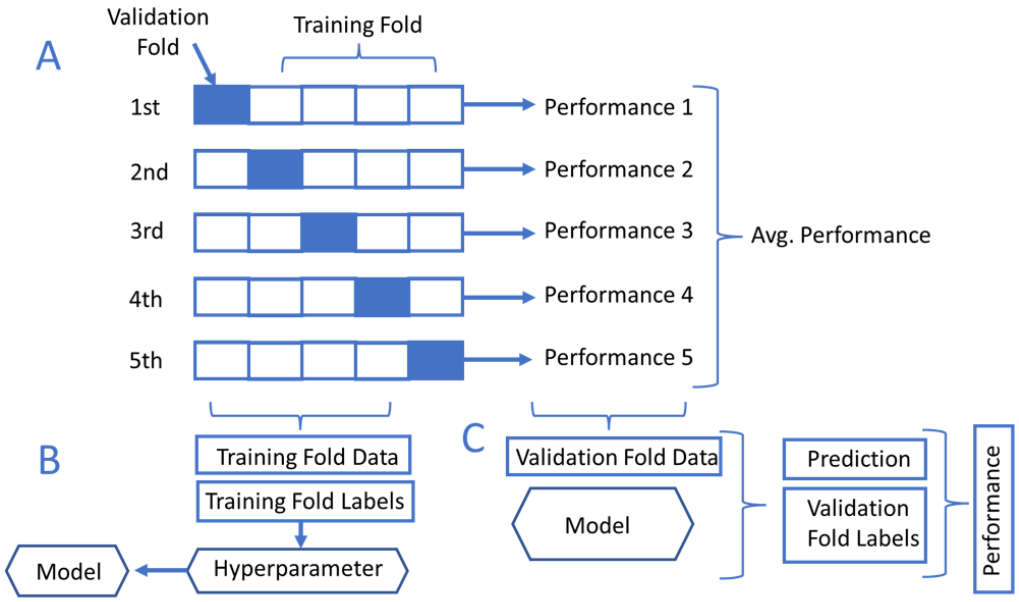


Figure 6.2: *K*-fold cross-validation process. (A) The *K* iterations of the cross-validation. (B) The training fold data and labels. (C) Evaluating model performance during the validation fold data stage.

6.1.4 Dice Score Coefficient (DC)

Dice's similarity coefficient is one of the most important measures to evaluate image segmentation techniques, especially in deep learning algorithms. The Dice coefficient is a statistical tool measuring the similarity between two data sets. The the coefficient is important especially in computer vision applications, as it can compare the segmented object to the ground truth data and give a sense of how accurate the algorithm is. Equation (6.4) shows the Dice score similarity coefficient formula.

$$Dice = 2x(\frac{U}{A}) \tag{6.4}$$

where (U) is the intersection area between two objects and (A) is the total area of two objects.

6.1.5 Mean Square Error (MSE)

The mean square error is the average squared error of the datasets. It is a typical evaluation metric to evaluate how accurate the predicted data are compared to the reference data. Equation (6.5) describes the mean square error mathematical formula.

$$MSE = \frac{1}{n}(\sum_{i=1}^{n}(x_i - y_i)^2) \tag{6.5}$$

where (x) is the predicted value, (y) is the observed value, and (n) is the number of data points.

6.2 Comparative Evaluation Results and Discussion

Qualitative and quantitative evaluations were used to evaluate the performance of the tongue segmentation from ultrasound images. Traditional and machine learning algorithms have different abilities for tongue image recognition to make each methodology unique on its own. In the qualitative analysis, we propose a qualitative scoring matrix that considers the final image quality, shape consistency, and algorithm complexity to test the method's usability performance. In the quantitative evaluation, we consider the MSD as a primary measure and some other measures such as the RMSE, MSE, and word error rate as secondary measures for some other applications.

6.2.1 Qualitative Evaluation

Among the traditional techniques that are based on the snake algorithm, the multihypothesis approach [15] produces robust research to handle tongue tracking efficiently. The output image quality is acceptable for speech recognition tasks. However, the quality of the image depends on the number of particle filters that are used, which makes this technique not practical for real-time applications. The algorithm is also tuned based on the tongue shape and motion model derived from different image frames. There is a trade-off for using a motion model. It may help to increase the confidence ratio of the segmented tongue contour. However, at the same time, the derived motion model may be inaccurate and cannot be applied in a general perspective. The research in [15] has some limitations that can be

addressed efficiently using deep learning algorithms based on an attention mechanism such as Transformer.

Publicly available tools such as EdgeTrack [13] and TongueTrack [14] are inefficient in real-time processing. They are susceptible to sudden and frequent failure during the segmentation and require a manual reinitialization to continue the processing. The image quality for their segmented contour is fair but is not suitable for medical-grade applications. These algorithms could not address the missing data issue and the variation of the shape consistency. The main drawback of these algorithms comes from the heavy optimization of too many parameters. The optimization issue does not just make them slow but also very limited to a specific subset of data and they cannot be applied for real challenges outside the lab. TongueTrack has an advantage over EdgeTrack by considering the spatial information between different frames. We believe if they used image denoising and a region-of-interest selection, the burden of computation complexity could be minimized. For future work suggestions, using a U-net architecture could be efficient for removing image noise and extracting image features, then combining them with existing algorithms as a hybrid technique.

The biomechanical method [110] derived a motion model for the tongue contour geometrical movement based on previously labelled X-ray images. The motion model alongside a Harris feature extractor were used to track the tongue features. The Harris feature extractor has too many limitations because it is sensitive to noise and requires localization constraints to select tongue contour features around the desired region of interest. In real-time tracking techniques, it may not be accurate since tongue motion may be more significant than the suggested local constraints. The final image and the extracted contour are susceptible to a high degree of uncertainty, making it not efficient for prediction using the suggested pipeline. The idea of using X-ray images to extract the motion model is good if we consider image quality compared to ultrasound. However, it could be risky to train the data from data with different distributions or statistical characteristics, requiring additional analysis. In future work, we recommend using deep learning algorithms instead of unrealistic motion models to merge ultrasound and X-ray images. Image fusion with deep learning models could be a potential solution for this problem as they can merge the quality of X-ray and ultrasound images using some image features or landmarks.

On the other hand, [8, 16], unlike most traditional techniques, implemented

denoising techniques to enhance the image and refine the tracking accuracy. However, the paper [16] relied on the snake algorithm as a base algorithm but with an automated reinitialization technique. The automatic reinitialization technique was robust enough to handle the sudden failure of the active contour. It might be more efficient than EdgeTrack and TongueTrack. However, the algorithm [16] still relied on too many constraints to optimize the snake algorithm. As mentioned before, this limits the ability to predict and estimate tongue displacement in a global context, making it unrealistic to predict the performance of any new data from a new source. In comparison, the research proposed in [8] went in a different direction to track the tongue without using the snake algorithm. A combined curvelet and shock filter denoised the image, then based on the temporal information of previous contours, an adaptive tongue region of interest was implemented. To extract a unique signature of each speaker, the tongue feature was extracted and transformed into speech time series data. In future research, we recommend combining the algorithm proposed in [8] with deep learning. The proposed research in [8] was robust for feature extraction using a policy-based adaptive model to extract features but had some limitations for real-time applications. Similarly, we recommend the algorithm [16] as a postprocessing tool combined with deep learning in a hybrid tongue contour extraction and refinement technique.

In deep learning methodologies, the research on convolutional neural networks to automate tongue segmentation [117] used the de facto segmentation models in biomedical imaging analysis, U-net and Dense U-Net. Dense U-Net had more generalization capability, meaning it could extract more features in a global context. It would be more accurate for any dataset outside the training set. However, Dense U-net is slower than the traditional U-net architecture which makes traditional U-net more efficient in real-time segmentation. Autotrace [111] used a translational deep belief network for image segmentation and was improved by [112] using a deep autoencoder. The deep autoencoder relied on the user data input, which affected the results for a limited context of given data. BowNet and wBowNet [119] and TongueNet [120] suggested two techniques for the tongue segmentation task based on multiscale contextual information and a deep network of landmarks. In general, most deep learning algorithms are based on CNNs, which is helpful for feature extraction and noise removal in a local context. However, the intrarelationship between the sequential image frames is limited. We suggest combining a CNN and any other deep learning-based spatial–temporal analysis to process con-

tinuous data. Some of the suggested algorithms are Vision Transformer, Vision-Graph, and ConvLSTM.

The authors in [123] proposed a ConvLSTM architecture. ConvLSTM is a novel approach that derives temporal information from the ultrasound images by extracting the intraframe relationship to resolve the issue of the lack of temporal resolution of other techniques. The model could predict tongue shape in the consecutive nine frames based on the data from the previous eight frames. In the same manner, [124] proposed a tongue contour tracking algorithm using a state-of-the-art U-net architecture alongside a temporal shape-consistency-based regularizer. This methodology was one of the most reliable techniques for real-time tongue processing. In their method, they used it to predict future frames, which could be used for training larger and more efficient algorithms such as the Transformer model. The Transformer model is gaining popularity as the state-of-the-art algorithm in the field due to its performance and predictability. The Transformer model also has some limitations, and it needs a huge dataset for training; this could be alleviated using the transfer learning methodology. Moreover, Transformer requires a fixed size of the input. LSTM also has limited memory but does not need a huge dataset like Transformer. The final suggestion is to use attention-based algorithms such as the Transformer model if the dataset is huge. If the dataset is small, LSTM can be used. Regarding image quality for deep learning, U-net is well known for preserving image features and noise removal. At the same time, attention-based algorithms are robust for predicting the correct speech behaviour to produce a high-quality output.

Figures 6.3 and 6.4 depict the quality evaluation matrix and bar chart for the total qualitative score of each category of tongue segmentation techniques. Image quality is generic and difficult to measure. Due to the lack of a definitive standard for image quality, we are proposing a new matrix that scores image quality based on different factors. In order to determine the image quality, we use the visual inspection and structural similarity index measure [268, 269]. In the usability measure, we mainly consider the algorithms' generalization and scalability. A generalized algorithm is one that performs well in real-life situations as well as in lab testing.

For the scalability measure, we define an algorithm as scalable if it is not sensitive to the variation in use-case environments or data size. This is crucial to ensure the algorithm is viable for use in different scenarios, not just optimized for one

solution. The consistency of shape is essential to determine whether the predicted shape is actually a tongue or not. We measure the shape consistency by comparing the results with ground-truth-labelled images and the data collected from different algorithms. The qualitative evaluation matrix is scored on a (0–5) scale (zero is the lowest and five is the highest). The final quality score is depicted on a percentile scale and evaluated with a satisfaction rate from low to high.

Category	Usability	Image Quality	Shape Consistency	Qualitative Score	Total Quality Satisfaction	Quality Score Scale (0-5)
Machine Learning (Overall)	3.50	3.50	4.00	73.33%	Medium	5
ML-based Spatial-Temporal	4.50	4.50	4.50	90.00%	High	4
						3
Traditional Techniques (Overall)	2.83	2.67	2.83	55.56%	Medium	2
Snake-based Algorithms	3.00	2.50	3.50	60.00%	Medium	1
Graph-based algorithms	3.50	3.00	3.00	63.33%	Medium	0
Biomechanical Method	2.00	2.50	2.00	43.33%	Low	

Figure 6.3: Quality evaluation matrix. Usability, image quality, and shape consistency are scored on a (0–5) scale (0 is the lowest and 5 is the highest). The final quality score is shown on a percentile scale and a satisfaction rate from low to high.

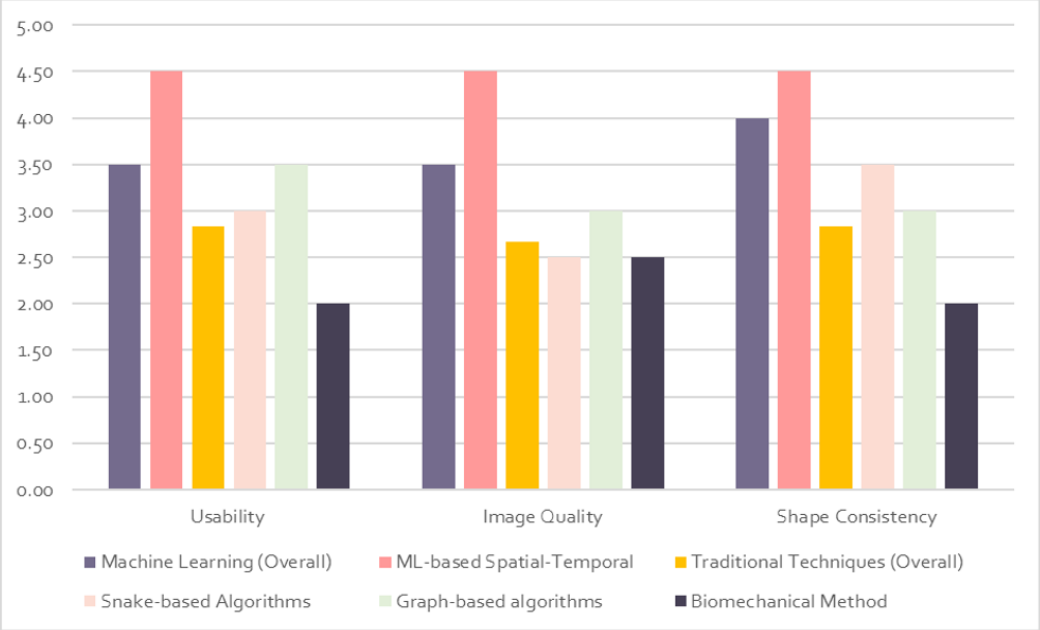


Figure 6.4: Bar chart for the total qualitative score of tongue image segmentation categories. The Y-axis is the qualitative score probability, and the X-axis is the quality score category for each image segmentation technique.

6.2.2 Quantitative Evaluation

The primary quantitative measure to evaluate tongue contour segmentation in this article was the MSD. The MSD is valid for this problem as it uses averaged measures to account for the tongue contour variation. The average MSD for the machine learning approaches was 1.4 mm, and the average MSD for the traditional techniques was 1.65 mm. The accuracy of these measures can be arguable as it is difficult to judge these results in realistic applications. These methods are never used in production and never tested outside the lab.

One of the common challenges in image recognition or machine learning is when the designed models typically fail when used outside the lab while they pass the testing stage in the lab. Poor performance may result from a small training dataset or an insufficiently generalized model (a generalized model performs well in testing and training). To transfer the model from research to the successful production stage, we recommend using a cloud-based solution to scale the designed model and evaluate the performance in different environments. In order to increase dataset diversity, we recommend data augmentation techniques. Moreover, transfer learning could be a viable solution if limited data are available.

Transfer learning is using features from pretrained models weights such as Imagenet [270] or VGG19 [271] and then fine-tuning the algorithm on the target datasets of the tongue images. Transfer learning minimizes the training time and enriches the model with low-level features such as edges and textures to help with data size limitation and to obtain more statistically accurate results. On the other hand, data augmentation helps to generate new data. Data augmentation can be simple, such as transforming data, rotating it, and flipping it, or more complex, such as creating new images using generative adversarial networks (GANs) [272].

There are different validation measures considered in addition to the MSD. Some of these measures are RMSE, MSE, speech recognition success ratio, word error rate, mean segmentation error, and accuracy. The fact is that there is no definitive recipe for the validation, and a combination of different measures is needed to address each methodology.

The MSD is considered a reasonable measure compared to the RMSE and MSE. For instance, the RMSE is helpful in regression analysis when we want to consider lower residual values unlike the MSE, which is biased towards higher values. The RMSE was used in [104] and the reported result was 0.2–0.3 mm, which was not

meaningful statistically to be considered as a reference for tongue segmentation standard. The MSE was reported in [123] and the the result was 17.3 mm. The better MSE is, the closer to zero. The problem with this measure is that it is sensitive to outliers or abnormal values, which maximize higher values; this explains why the error was high in [123]. To use the MSE correctly, the researcher should be careful in the feature engineering stage to remove unnecessary data. A logarithmic scale sometimes helps in this case. Accuracy was also used in the biomechanical method [110]; they reported a result of 0.62–0.97 mm. Accuracy is a generic and simple evaluation measure. It has severe limitations in the case of data imbalance and does not account for the variation in data size.

Some other used measures such as speech recognition success ratio which was reported in [126] as 65% for their algorithm evaluation. It only provides a counting measure for the final speech success rate, but not for the tongue segmentation accuracy. It is not valid in the case of data variation, since it neither considers nor accounts for the statistical distribution. The word error rate was also reported in [122]. It can provide a general impression of performance, but it does not provide any meaningful or accurate information about the tongue; it does not provide any clinical measure.

The mean segmentation error was used in [109]; their results were reported for dense and sparse data as 4.49 mm and 2.23 mm, respectively. This technique was compared to the MSE, but the researchers enhanced it by adding additional optimization techniques to remove unnecessary data. This is a significant enhancement compared to the MSE evaluations, but it is not as efficient as the MSD, which represents the most reasonable measure that can be valid to evaluate tongue segmentation techniques.

Table 6.1 compares the most important techniques used to segment tongue contour from ultrasound images by describing each method's core methodologies, results, data types, and limitations.

Table 6.1: Comparison of tongue contour segmentation methodologies.

Method	Category	EV. Measure	EV. Result	Data Type	Core Methodologies
EdgeTrack [13]	Traditional	MSD	0.53-1.0mm	US	Snake algorithm + gradient+ local image information, and object orientation.
TongueTrack [14]	Traditional	MSD	3mm	US	Higher-order Markov random field energy minimization framework.
Tongue shape prediction from landmarks [104]	Traditional	RMSE	(0.2-0.3) mm	US+ EMA or X-ray	Spline interpolation + Landmark mapping using metal pellets.
Graph-based [109]	Traditional	Mean Segmentation Error	Dense=4.49 mm Sparse=2.23mm	US	Image graph-based analysis+ Adaptive temporal regularization using Markov random field optimization.
Biomechanical [110]	Traditional	Accuracy	0.62mm-0.97mm.	X-ray and US	Harries features + Optical flow.
Multi-Hypothesis Approach [15]	Traditional and ML	MSD	1.69 ±1.10 mm	US	Snake algorithm + particle filter.
Computer Vision-based Tongue Tracking and Feature Extraction [8]	Traditional	MSD	0.933mm	US	Image Denoising + tongue adaptive localization + feature extraction + data transformation and analysis.
Fully automate the tongue contour extraction [16]	Traditional	MSD	1.01mm - 0.63mm	US	Snake algorithm + phase symmetry filter + algorithm resetting.
Autotrace [111]	ML	MSD	0.73mm	US	Deep Learning + translational deep belief network.

Continue for **Table** 6.1.

Method	Category	EV. Measure	EV. Result	Data Type	Core Methodologies
Enhanced Autotrace [112]	ML	MSD	1.0mm	US	Deep Auto-Encoder.
CNN to automate the tongue segmentation [117]	ML	MSD	U-net=5.81mm. (Dense U-net)=5.6mm.	US	U-net + Dense U-net.
Bownet and Bownet [119]	ML	MSD	1.4mm	US	Deep Network in landmarks.
TongueNet [120]	ML	MSD	0.31 pixel	US	Multi-scale contextual information + Dilated convolution.
DCAE-based B-Mode US [122]	ML	Word Error Rate	6.17 %	US	Denoising Convolutional Autoencoder (DCAE).
ConvLSTM [123]	ML	MSE and (CW-SSIM)	MSE = 17.13. CWSSIM = 0.932.	Tongue US images.	(CNN) + (LSTM).
U-NET and Shape Consistency-based Regularizer [124]	Traditional and ML	MSD	(2.243±0.026) mm	US	Unet architecture + Temporal continuity using shape consistency-based regularizer.
wUnet [125]	ML	MSD	1.18mm	US	Unet architecture + VGG19 block instead of skip connections.
SottoVoce [126]	ML	Speech Recognition Success Ratio	65%	US + Speech audio recording.	Deep CNN.

Chapter 7

Conclusion and Future Work

Conclusion

The dissertation discusses the usability and importance of ViT-based artificial intelligence architecture in medical imaging segmentation with a case study on lingual ultrasound applications. ViT and UNet were utilized to build TongueTransUNet architecture and a dynamic quality control process was added to manage the input source and output results. In addition, the quality control process considers features in a latent space and verifies the shape consistency alongside the accuracy measures of the extracted object.

The research also extended to manage the access control for the data journey within the telehealth and cloud-AI infrastructure. In this research, a zero-trust context-aware system was harnessed to evaluate the similarity, data integrity, and consistency between attributes of input source, users, and devices to score the event at each data transaction. This step is essential to minimize medical errors and manage access to secure the telehealth information system.

The research outcomes and contributions are summarized in the following items:

- Build TongueTransUNet which is used to manage quality control of features extraction and processing. The process harness latent space and automated chain of trusted events, in addition to the human-reinforcement feedback process.
- Develop a ZTCloudGuard framework to maintain a secure, confidential and available data exchange between devices,users and output.

- Consider and extract the hardware fingerprint from the overall system, human identity verification and output data content and pattern.
- Propose a combined evaluation metrics using quantitative and qualitative techniques to assess the final output.

Future Work

The current research focuses mainly on managing the quality control of data ingestion and processing to enhance AI safety and consider consistency, integrity and availability within distributed and complex cloud-based system. The case study is in the healthcare information system. There are few recommendations to improve the research presented in the dissertation. Below there are few itemized recommendations that may improve the research.

- Combine the current system with other modalities to utilize multi-sensors in a multi-dimensional space. For example, utilizing ultrasound image fusion with different imaging modalities like MRI, CT and X-ray in real-time applications can be useful to reduce the cost, time and radiation dose.
- Redesign large language models by considering the current proposed research and build the current system cumulatively in a sequential basis. Each data transaction includes a series of events in a multi-clustered subspaces. The chain of events should be trusted and relevant to each other. This could help to generate subject-relevant medical reports.
- Design a multi-layered distributed and encrypted blockchain access control system. This system can be useful for cloud-based system to ensure each entity within cloud-AI environment is safe and no one can access data for different users.

References

[1] F. Gaillard, "Muscles of the tongue: Radiology reference article," May 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[3] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," arXiv preprint arXiv:2201.09873, 2022.

[4] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," Diagnostics, vol. 11, no. 8, p. 1384, 2021.

[5] J. Zhang, Y. Nie, J. Chang, and J. J. Zhang, "Surgical instruction generation with transformers," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pp. 290–299, Springer International Publishing, 2021.

[6] C. Meng, L. Trinh, N. Xu, and Y. Liu, "MIMIC-IF: Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset," apr 2021.

[7] K. Al-Hammuri, Computer vision-based tracking and feature extraction for lingual ultrasound. PhD thesis, 2019.

[8] K. Al-hammuri, Computer Vision-based tracking and feature extraction for lingual ultrasound. PhD thesis, University of Victoria, 2019.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International conference on machine learning, pp. 1597–1607, PMLR, 2020.

[10] K. Al-hammuri, F. Gebali, I. Thirumarai Chelvan, and A. Kanan, "Tongue contour tracking and segmentation in lingual ultrasound for speech recognition: A review," Diagnostics, vol. 12, no. 11, p. 2811, 2022.

[11] K. Al-hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," Visual Computing for Industry, Biomedicine, and Art, vol. 6, jul 2023.

[12] K. Al-Hammuri, F. Gebali, and A. Kanan, "Ztcloudguard: Zero trust context-aware access management framework to avoid medical errors in the era of generative ai and cloud-based health information ecosystems," AI, vol. 5, no. 3, pp. 1111–1131, 2024.

[13] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," Clinical linguistics & phonetics, vol. 19, no. 6-7, pp. 545–554, 2005.

[14] L. Tang, T. Bressmann, and G. Hamarneh, "Tongue contour tracking in dynamic ultrasound via higher-order mrfs and efficient fusion moves," Medical image analysis, vol. 16, no. 8, pp. 1503–1520, 2012.

[15] C. Laporte and L. Ménard, "Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech," Medical image analysis, vol. 44, pp. 98–114, 2018.

[16] E. Karimi, L. Ménard, and C. Laporte, "Fully-automated tongue detection in ultrasound images," Computers in Biology and Medicine, vol. 111, p. 103335, 2019.

[17] J. Cai, B. Denby, P. Roussel-Ragot, G. Dreyfus, and L. Crevier-Buchman, "Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model.," in Interspeech, pp. 1005–1008, 2011.

[18] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal sensors and deep learning-based speech recognition: A review," Sensors, vol. 21, no. 4, p. 1399, 2021.

[19] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, "Silent versus modal multi-speaker speech recognition from ultrasound and video," arXiv preprint arXiv:2103.00333, 2021.

[20] M. Stone, "A guide to analysing tongue motion from ultrasound images," Clinical linguistics & phonetics, vol. 19, no. 6-7, pp. 455–501, 2005.

[21] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time mri," Computer Speech & Language, vol. 52, pp. 1–22, 2018.

[22] M. Deng, D. Leotta, G. Huang, Z. Zhao, and Z. Liu, "Craniofacial, tongue, and speech characteristics in anterior open bite patients of east african ethnicity," Res Rep Oral Maxillofac Surg, vol. 3, no. 1, p. 21, 2019.

[23] S. G. Lingala, A. Toutios, J. Töger, Y. Lim, Y. Zhu, Y.-C. Kim, C. Vaz, S. S. Narayanan, and K. S. Nayak, "State-of-the-art mri protocol for comprehensive assessment of vocal tract structure and function.," in Interspeech, pp. 475–479, 2016.

[24] Ö. D. Köse and M. Saraçlar, "Multimodal representations for synchronized speech and real-time mri video processing," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1912–1924, 2021.

[25] K. Isaieva, Y. Laprie, A. Houssard, J. Felblinger, and P.-A. Vuissoz, "Tracking the tongue contours in rt-mri films with an autoencoder dnn approach," in ISSP 2020-12th International Seminar on Speech Production, 2020.

[26] Z. Zhao, Y. Lim, D. Byrd, S. Narayanan, and K. S. Nayak, "Improved 3d real-time mri of speech production," Magnetic Resonance in Medicine, vol. 85, no. 6, pp. 3182–3195, 2021.

[27] F. Xing, Three Dimensional Tissue Motion Analysis from Tagged Magnetic Resonance Imaging. PhD thesis, Johns Hopkins University, 2015.

[28] F. Höwing, L. S. Dooley, and D. Wermser, "Tracking of non-rigid articulatory organs in x-ray image sequences," Computerized medical imaging and graphics, vol. 23, no. 2, pp. 59–67, 1999.

[29] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Ferbach-Hecker, L. Ma, et al., "An x-ray database, tools and procedures for the study of speech production," in ISSP 2011-9th International Seminar on Speech Production, pp. 41–48, 2011.

[30] J. Yu, "Speech synchronized tongue animation by combining physiology modeling and x-ray image fitting," in International Conference on Multimedia Modeling, pp. 726–737, Springer, 2017.

[31] R. Guijarro-Martínez and G. Swennen, "Cone-beam computerized tomography imaging and analysis of the upper airway: a systematic review of the literature," International journal of oral and maxillofacial surgery, vol. 40, no. 11, pp. 1227–1237, 2011.

[32] T.-N. Hou, L.-N. Zhou, and H.-J. Hu, "Computed tomographic angiography study of the relationship between the lingual artery and lingual markers in patients with obstructive sleep apnoea," Clinical radiology, vol. 66, no. 6, pp. 526–529, 2011.

[33] S.-H. Kim and S.-K. Choi, "Changes in the hyoid bone, tongue, and oropharyngeal airway space after mandibular setback surgery evaluated by cone-beam computed tomography," Maxillofacial Plastic and Reconstructive Surgery, vol. 42, no. 1, pp. 1–9, 2020.

[34] A. Sierhej, J. Verhoeven, N. R. Miller, and C. C. Reyes-Aldasoro, "Optimisation strategies for the registration of computed tomography images of electropalatography," bioRxiv, 2020.

[35] X. Guo, X. Liang, J. Jin, J. Chen, J. Liu, Y. Qiao, J. Cheng, and J. Zhao, "Three-dimensional computed tomography mapping of 136 tongue-type calcaneal fractures from a single centre," Annals of Translational Medicine, vol. 9, no. 24, 2021.

[36] M. Yang, J. Tao, and D. Zhang, "Extraction of tongue contour in x-ray videos," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1094–1098, IEEE, 2013.

[37] C. Luo, R. Li, L. Yu, J. Yu, and Z. Wang, "Automatic tongue tracking in x-ray images," Chinese Journal of Electronics, vol. 24, no. 4, pp. 767–771, 2015.

[38] Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch, "Articulatory copy synthesis from cine x-ray films," in InterSpeech-14th Annual Conference of the International Speech Communication Association-2013, 2013.

[39] M.-O. Berger, G. erard Mozelle, and Y. Laprie, "Cooperation of active contours and optical ow for tongue tracking in x-ray motion pictures," 1995.

[40] G. Thimm, "Tracking articulators in x-ray movies of the vocal tract," in International Conference on Computer Analysis of Images and Patterns, pp. 126–133, Springer, 1999.

[41] A. Koren, L. D. Grošelj, and I. Fajdiga, "Ct comparison of primary snoring and obstructive sleep apnea syndrome: role of pharyngeal narrowing ratio and soft palate-tongue contact in awake patient," European archives of oto-rhino-laryngology, vol. 266, no. 5, pp. 727–734, 2009.

[42] T. Uysal, A. Yagci, F. I. Ucar, I. Veli, and T. Ozer, "Cone-beam computed tomography evaluation of relationship between tongue volume and lower incisor irregularity," The European Journal of Orthodontics, vol. 35, no. 5, pp. 555–562, 2013.

[43] Y. Shigeta, T. Ogawa, E. Ando, G. T. Clark, and R. Enciso, "Influence of tongue/mandible volume ratio on oropharyngeal airway in japanese male patients with obstructive sleep apnea," Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, vol. 111, no. 2, pp. 239–243, 2011.

[44] X. Ding, S. Suzuki, M. Shiga, N. Ohbayashi, T. Kurabayashi, and K. Moriyama, "Evaluation of tongue volume and oral cavity capacity using cone-beam computed tomography," Odontology, vol. 106, no. 3, pp. 266–273, 2018.

[45] S. Rana, O. Kharbanda, and B. Agarwal, "Influence of tongue volume, oral cavity volume and their ratio on upper airway: A cone beam computed tomography study," Journal of Oral Biology and Craniofacial Research, vol. 10, no. 2, pp. 110–117, 2020.

[46] G. Eggers, B. Kress, S. Rohde, and J. Muhling, "Intraoperative computed tomography and automated registration for image-guided cranial surgery," Dentomaxillofacial Radiology, vol. 38, no. 1, pp. 28–33, 2009.

[47] W. P. Liu, J. D. Richmon, J. M. Sorger, M. Azizian, and R. H. Taylor, "Augmented reality and cone beam ct guidance for transoral robotic surgery," Journal of robotic surgery, vol. 9, no. 3, pp. 223–233, 2015.

[48] Y.-W. Zhong, Y. Jiang, S. Dong, W.-J. Wu, L.-X. Wang, J. Zhang, and M.-W. Huang, "Tumor radiomics signature for artificial neural network-assisted detection of neck metastasis in patient with tongue cancer," Journal of Neuroradiology, vol. 49, no. 2, pp. 213–218, 2022.

[49] S. Khanal, M. T. Johnson, and N. Bozorg, "Articulatory comparison of l1 and l2 speech for mispronunciation diagnosis," in 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 693–697, IEEE, 2021.

[50] S. Medina, D. Tome, C. Stoll, M. Tiede, K. Munhall, A. G. Hauptmann, and I. Matthews, "Speech driven tongue animation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20406–20416, 2022.

[51] J. A. Shaw, S. Oh, K. Durvasula, and A. Kochetov, "Articulatory coordination distinguishes complex segments from segment sequences," Phonology, vol. 38, no. 3, pp. 437–477, 2021.

[52] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," Speech Communication, vol. 55, no. 1, pp. 22–32, 2013.

[53] L. A. Cheah, J. M. Gilbert, J. A. Gonzalez, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "Towards an intraoral-based silent speech restoration system for post-laryngectomy voice replacement," in International Joint Conference on Biomedical Engineering Systems and Technologies, pp. 22–38, Springer, 2016.

[54] J. A. Gonzalez and P. D. Green, "A real-time silent speech system for voice restoration after total laryngectomy," Revista de logopedia, foniatría y audiología, vol. 38, no. 4, pp. 148–154, 2018.

[55] L. A. Cheah, J. M. Gilbert, J. A. González, P. D. Green, S. R. Ell, R. K. Moore, and E. Holdsworth, "A wearable silent speech interface based on magnetic sensors with motion-artefact removal.," in BIODEVICES, pp. 56–62, 2018.

[56] N. Sebkhi, A novel wireless tongue tracking system for speech applications. PhD thesis, Georgia Institute of Technology, 2019.

[57] A. Lee, M. Liker, Y. Fujiwara, I. Yamamoto, Y. Takei, and F. Gibbon, "Epg research and therapy: further developments," Clinical Linguistics & Phonetics, pp. 1–21, 2022.

[58] L.-C. Chen, P.-H. Chen, R. T.-H. Tsai, and Y. Tsao, "Epg2s: Speech generation and speech enhancement based on electropalatography and audio signals using multimodal learning," IEEE Signal Processing Letters, 2022.

[59] M. Wand, T. Schultz, and J. Schmidhuber, "Domain-adversarial training for session independent emg-based speech recognition.," in Interspeech, pp. 3167–3171, 2018.

[60] A. Ratnovsky, S. Malayev, S. Ratnovsky, S. Naftali, and N. Rabin, "Emg-based speech recognition using dimensionality reduction methods," Journal of Ambient Intelligence and Humanized Computing, pp. 1–11, 2021.

[61] H.-S. Cha, W.-D. Chang, and C.-H. Im, "Deep-learning-based real-time silent speech recognition using facial electromyogram recorded around eyes for hands-free interfacing in a virtual reality environment," Virtual Reality, pp. 1–11, 2022.

[62] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for emg-based human-machine interaction: A review," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 3, pp. 512–533, 2021.

[63] H. Hayashi and T. Tsuji, "Human–machine interfaces based on bioelectric signals: A narrative review with a novel system proposal," IEEJ Transactions on Electrical and Electronic Engineering, 2022.

[64] R. Harada, N. Hojyo, K. Fujimoto, and T. Oyama, "Development of communication system from emg of suprahyoid muscles using deep learning," in 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), pp. 5–9, IEEE, 2022.

[65] Q. Zhang, J. Jing, D. Wang, and R. Zhao, "Wearsign: Pushing the limit of sign language translation using inertial and emg wearables," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 6, no. 1, pp. 1–27, 2022.

[66] G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik, "Improving eeg based continuous speech recognition," arXiv preprint arXiv:1911.11610, 2019.

[67] K. GÖRÜR, M. R. BOZKURT, M. S. BASCIL, and F. TEMURTAS, "Tongue-operated biosignal over eeg and processing with decision tree and knn," Academic Platform-Journal of Engineering and Science, vol. 9, no. 1, pp. 112–125, 2021.

[68] M. Rao et al., "Decoding imagined speech using wearable eeg headset for a single subject," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2622–2627, IEEE, 2021.

[69] M. A. Bakhshali, M. Khademi, and A. Ebrahimi-Moghadam, "Investigating the neural correlates of imagined speech: An eeg-based connectivity analysis," Digital Signal Processing, vol. 123, p. 103435, 2022.

[70] M. Koctúrová and J. Juhár, "A novel approach to eeg speech activity detection with visual stimuli and mobile bci," Applied Sciences, vol. 11, no. 2, p. 674, 2021.

[71] H. Lovenia, H. Tanaka, S. Sakti, A. Purwarianti, and S. Nakamura, "Speech artifact removal from eeg recordings of spoken word production with tensor decomposition," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1115–1119, IEEE, 2019.

[72] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech recognition with no speech or with noisy speech," in ICASSP 2019-2019 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1090–1094, IEEE, 2019.

[73] Y.-E. Lee and S.-H. Lee, "Eeg-transformer: Self-attention from transformer architecture for decoding eeg of imagined speech," in 2022 10th International Winter Conference on Brain-Computer Interface (BCI), pp. 1–4, IEEE, 2022.

[74] G. Krishna, C. Tran, M. Carnahan, and A. Tewfik, "Improving eeg based continuous speech recognition using gan," arXiv preprint arXiv:2006.01260, 2020.

[75] I. Wilson, "Using ultrasound for teaching and researching articulation," Acoustical Science and Technology, vol. 35, no. 6, pp. 285–289, 2014.

[76] B. Gick, B. Bernhardt, P. Bacsfalvi, I. Wilson, and M. Zampini, "Ultrasound imaging applications in second language acquisition," Phonology and second language acquisition, vol. 36, no. 6, pp. 309–322, 2008.

[77] S. R. Li, S. Dugan, J. Masterson, H. Hudepohl, C. Annand, C. Spencer, R. Seward, M. A. Riley, S. Boyce, and T. D. Mast, "Classification of accurate and misarticulated /*a*r/for ultrasound biofeedback using tongue part displacement trajectories," Clinical Linguistics & Phonetics, pp. 1–27, 2022.

[78] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," arXiv preprint arXiv:1907.00835, 2019.

[79] L. McKeever, J. Cleland, and J. Delafield-Butt, "Using ultrasound tongue imaging to analyse maximum performance tasks in children with autism: a pilot study," Clinical Linguistics & Phonetics, vol. 36, no. 2-3, pp. 127–145, 2022.

[80] M. Castillo, F. Rubio, D. Porras, S. H. Contreras-Ortiz, and A. Sepúlveda, "A small vocabulary database of ultrasound image sequences of vocal tract dynamics," in 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1–5, IEEE, 2019.

[81] M. Ohkubo and J. M. Scobbie, "Tongue shape dynamics in swallowing using sagittal ultrasound," Dysphagia, vol. 34, no. 1, pp. 112–118, 2019.

[82] S. Chen, Y. Zheng, C. Wu, G. Sheng, P. Roussel, and B. Denby, "Direct, near real time animation of a 3d tongue model using non-invasive ultrasound images," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4994–4998, IEEE, 2018.

[83] Y. Ji, L. Liu, H. Wang, Z. Liu, Z. Niu, and B. Denby, "Updating the silent speech challenge benchmark with deep learning," Speech Communication, vol. 98, pp. 42–50, 2018.

[84] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," Speech Communication, vol. 52, no. 4, pp. 270–287, 2010.

[85] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," IEEE access, vol. 8, pp. 177995–178021, 2020.

[86] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," IEEE Transactions on medical imaging, vol. 25, no. 8, pp. 987–1010, 2006.

[87] H. Huang, Z. Ge, H. Wang, J. Wu, C. Hu, N. Li, X. Wu, and C. Pan, "Segmentation of echocardiography based on deep learning model," Electronics, vol. 11, no. 11, p. 1714, 2022.

[88] Y. Hu, Y. Guo, Y. Wang, J. Yu, J. Li, S. Zhou, and C. Chang, "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model," Medical physics, vol. 46, no. 1, pp. 215–228, 2019.

[89] T. Wang, Y. Lei, M. Axente, J. Yao, J. Lin, J. D. Bradley, T. Liu, D. Xu, and X. Yang, "Automatic breast ultrasound tumor segmentation via one-stage hierarchical target activation network," in Medical Imaging 2022: Ultrasonic Imaging and Tomography, vol. 12038, pp. 137–142, SPIE, 2022.

[90] Y. Lei, X. He, J. Yao, T. Wang, L. Wang, W. Li, W. J. Curran, T. Liu, D. Xu, and X. Yang, "Breast tumor segmentation in 3d automatic breast ultrasound using mask scoring r-cnn," Medical physics, vol. 48, no. 1, pp. 204–214, 2021.

[91] J. Yang, L. Tong, M. Faraji, and A. Basu, "Ivus-net: An intravascular ultrasound segmentation network," in International Conference on Smart Multimedia, pp. 367–377, Springer, 2018.

[92] H. Du, L. Ling, W. Yu, P. Wu, Y. Yang, M. Chu, J. Yang, W. Yang, and S. Tu, "Convolutional networks for the segmentation of intravascular ultrasound images: Evaluation on a multicenter dataset," Computer Methods and Programs in Biomedicine, vol. 215, p. 106599, 2022.

[93] M. B. Allan, M. H. Jafari, N. V. Woudenberg, O. Frenkel, D. Murphy, T. Wee, R. D'Ortenzio, Y. Wu, J. Roberts, N. Shatani, et al., "Multi-task deep learning for segmentation and landmark detection in obstetric sonography," in Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 12034, pp. 160–165, SPIE, 2022.

[94] S. N. Bushra and G. Shobana, "Obstetrics and gynaecology ultrasound image analysis towards cryptic pregnancy using deep learning-a review," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 949–953, IEEE, 2021.

[95] Z. Zhang and Y. Han, "Detection of ovarian tumors in obstetric ultrasound imaging using logistic regression classifier with an advanced machine learning approach," IEEE Access, vol. 8, pp. 44999–45008, 2020.

[96] T. G. Csapó, K. Xu, A. Deme, T. E. Gráczi, and A. Markó, "Transducer misalignment in ultrasound tongue imaging," in Proceedings of the 12th International Seminar on Speech Production, pp. 166–169, 2021.

[97] M. Stone and T. H. Shawker, "An ultrasound examination of tongue movement during swallowing," Dysphagia, vol. 1, no. 2, pp. 78–83, 1986.

[98] T. Kaburagi and M. Honda, "An ultrasonic method for monitoring tongue shape and the position of a fixed point on the tongue surface," The Journal of the Acoustical Society of America, vol. 95, no. 4, pp. 2268–2270, 1994.

[99] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," International journal of computer vision, vol. 1, no. 4, pp. 321–331, 1988.

[100] K. Iskarous, "Detecting the edge of the tongue: A tutorial," Clinical linguistics & phonetics, vol. 19, no. 6-7, pp. 555–565, 2005.

[101] Y. S. Akgul, C. Kambhamettu, and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," in Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231), pp. 298–303, IEEE, 1998.

[102] Y. S. Akgul, C. Kambhamettu, and M. Stone, "Automatic motion analysis of the tongue surface from ultrasound image sequences," in Proceedings. Workshop on Biomedical Image Analysis (Cat. No. 98EX162), pp. 126–132, IEEE, 1998.

[103] Y. S. Akgul, C. Kambhamettu, and M. Stone, "Automatic extraction and tracking of the tongue contours," IEEE Transactions on Medical Imaging, vol. 18, no. 10, pp. 1035–1045, 1999.

[104] C. Qin, M. A. Carreira-Perpinán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," 2008.

[105] K. Xu, Y. Yang, M. Stone, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel, and B. Denby, "Robust contour tracking in ultrasound tongue image sequences," Clinical linguistics & phonetics, vol. 30, no. 3-5, pp. 313–327, 2016.

[106] K. Xu, T. Gábor Csapó, P. Roussel, and B. Denby, "A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization," The Journal of the Acoustical Society of America, vol. 139, no. 5, pp. EL154–EL160, 2016.

[107] A. Roussos, A. Katsamanis, and P. Maragos, "Tongue tracking in ultrasound images with active appearance models," in 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 1733–1736, IEEE, 2009.

[108] M. Aron, A. Roussos, M.-O. Berger, E. Kerrien, and P. Maragos, "Multimodality acquisition of articulatory data and processing," in 2008 16th European Signal Processing Conference, pp. 1–5, IEEE, 2008.

[109] L. Tang and G. Hamarneh, "Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 154–161, IEEE, 2010.

[110] M. Loosvelt, P.-F. Villard, and M.-O. Berger, "Using a biomechanical model for tongue tracking in ultrasound images," in International Symposium on Biomedical Simulation, pp. 67–75, Springer, 2014.

[111] I. Fasel and J. Berry, "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech," in 2010 20th International Conference on Pattern Recognition, pp. 1493–1496, IEEE, 2010.

[112] A. Jaumard-Hakoun, K. Xu, P. Roussel-Ragot, G. Dreyfus, and B. Denby, "Tongue contour extraction from ultrasound images based on deep neural network," arXiv preprint arXiv:1605.05912, 2016.

[113] D. Fabre, T. Hueber, F. Bocquelet, and P. Badin, "Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks," in Interspeech 2015-16th Annual Conference of the International Speech Communication Association, 2015.

[114] K. Xu, P. Roussel, T. G. Csapó, and B. Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using b-mode ultrasound images," The Journal of the Acoustical Society of America, vol. 141, no. 6, pp. EL531–EL537, 2017.

[115] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, pp. 234–241, Springer, 2015.

[116] J. Zhu, W. Styler, and I. C. Calloway, "Automatic tongue contour extraction in ultrasound images with convolutional neural networks," The Journal of the Acoustical Society of America, vol. 143, no. 3, pp. 1966–1966, 2018.

[117] J. Zhu, W. Styler, and I. Calloway, "A cnn-based tool for automatic tongue contour tracking in ultrasound images," arXiv preprint arXiv:1907.10210, 2019.

[118] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708, 2017.

[119] M. H. Mozaffari and W.-S. Lee, "Encoder-decoder cnn models for automatic tracking of tongue contours in real-time ultrasound data," Methods, vol. 179, pp. 26–36, 2020.

[120] M. H. Mozaffari, N. Yamane, and W.-S. Lee, "Deep learning for automatic tracking of tongue surface in real-time ultrasound videos, landmarks instead of contours," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2785–2792, IEEE, 2020.

[121] S. Wen, Automatic tongue contour segmentation using deep learning. PhD thesis, University of Ottawa, 2018.

[122] B. Li, K. Xu, D. Feng, H. Mi, H. Wang, and J. Zhu, "Denoising convolutional autoencoder based b-mode ultrasound tongue image feature extraction," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7130–7134, IEEE, 2019.

[123] C. Zhao, P. Zhang, J. Zhu, C. Wu, H. Wang, and K. Xu, "Predicting tongue motion in unlabeled ultrasound videos using convolutional lstm neural networks," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5926–5930, IEEE, 2019.

[124] M. Feng, Y. Wang, K. Xu, H. Wang, and B. Ding, "Improving ultrasound tongue contour extraction using u-net and shape consistency-based regularizer," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6443–6447, IEEE, 2021.

[125] G. Li, J. Chen, Y. Liu, and J. Wei, "wunet: A new network used for ultrasonic tongue contour extraction," Speech Communication, 2022.

[126] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–11, 2019.

[127] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in IEEE 2011 workshop on automatic speech recognition and understanding, no. CONF, IEEE Signal Processing Society, 2011.

[128] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in 2008 eighth ieee international conference on data mining, pp. 413–422, IEEE, 2008.

[129] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," arXiv preprint arXiv:2202.10108, 2022.

[130] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, 2022.

[131] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," IET Image Processing, vol. 16, no. 5, pp. 1243–1267, 2022.

[132] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in International conference on medical image computing and computer-assisted intervention, pp. 586–594, Springer, 2018.

[133] Y. Wang, H. Xie, S. Fang, M. Xing, J. Wang, S. Zhu, and Y. Zhang, "Petr: Rethinking the capability of transformer-based language model in scene text recognition," IEEE Transactions on Image Processing, vol. 31, pp. 5585–5598, 2022.

[134] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[135] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[136] Y. Gao, J. M. Phillips, Y. Zheng, R. Min, P. T. Fletcher, and G. Gerig, "Fully convolutional structured lstm networks for joint 4d medical image segmentation," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1104–1108, IEEE, 2018.

[137] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[138] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," IEEE Transactions on Instrumentation and Measurement, 2022.

[139] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[140] Y. Maeda, N. Fukushima, and H. Matsuo, "Taxonomy of vectorization patterns of programming for fir image filters using kernel subsampling and new one," Applied Sciences, vol. 8, no. 8, p. 1235, 2018.

[141] P. Jain, S. Vijayanarasimhan, and K. Grauman, "Hashing hyperplane queries to near points with applications to large-scale active learning," Advances in Neural Information Processing Systems, vol. 23, 2010.

[142] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," Neural computation, vol. 31, no. 7, pp. 1235–1270, 2019.

[143] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[144] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in International conference on machine learning, pp. 1243–1252, PMLR, 2017.

[145] S. Takase, S. Kiyono, S. Kobayashi, and J. Suzuki, "On layer normalizations and residual connections in transformers," arXiv preprint arXiv:2206.00330, 2022.

[146] M. O. Topal, A. Bas, and I. van Heerden, "Exploring transformers in natural language generation: Gpt, bert, and xlnet," arXiv preprint arXiv:2102.08036, 2021.

[147] S. Wang, F. Liu, and B. Liu, "Escaping the gradient vanishing: Periodic alternatives of softmax in attention mechanism," IEEE Access, vol. 9, pp. 168749–168759, 2021.

[148] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.

[149] H. Taud and J. Mas, "Multilayer perceptron (mlp)," in Geomatic approaches for modeling land change scenarios, pp. 451–455, Springer, 2018.

[150] A. A. Akinyelu, F. Zaccagna, J. T. Grist, M. Castelli, and L. Rundo, "Brain tumor diagnosis using machine learning, convolutional neural networks, capsule neural networks and vision transformers, applied to mri: A survey," Journal of Imaging, vol. 8, no. 8, p. 205, 2022.

[151] E. Mahoro and M. A. Akhloufi, "Breast cancer classification on thermograms using deep cnn and transformers," Quantitative InfraRed Thermography Journal, pp. 1–20, 2022.

[152] A. Shmatko, N. Ghaffari Laleh, M. Gerstung, and J. N. Kather, "Artificial intelligence in histopathology: enhancing cancer research and clinical oncology," Nature Cancer, vol. 3, no. 9, pp. 1026–1038, 2022.

[153] C. McMaster, A. Bird, D. F. Liew, R. R. Buchanan, C. E. Owen, W. W. Chapman, and D. E. Pires, "Artificial intelligence and deep learning for rheumatologists," Arthritis & Rheumatology, 2022.

[154] D.-R. Beddiar, M. Oussalah, and T. Seppänen, "Automatic captioning for medical imaging (mic): a rapid review of literature," Artificial Intelligence Review, pp. 1–58, 2022.

[155] F. Renna, M. Martins, A. Neto, A. Cunha, D. Libânio, M. Dinis-Ribeiro, and M. Coimbra, "Artificial intelligence for upper gastrointestinal endoscopy: A roadmap from technology development to clinical practice," Diagnostics, vol. 12, no. 5, p. 1278, 2022.

[156] L. Coan, B. Williams, M. K. A. Venkatesh, S. Upadhyaya, A. Al Kafri, S. Czanner, R. Venkatesh, C. E. Willoughby, S. Kavitha, and G. Czanner, "Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review," Survey of ophthalmology, 2022.

[157] A. Chang, "The role of artificial intelligence in digital health," in Digital health entrepreneurship, pp. 71–81, Springer, 2020.

[158] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," arXiv preprint arXiv:2105.05537, 2021.

[159] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using u-net based fully convolutional networks," in annual conference on medical image understanding and analysis, pp. 506–517, Springer, 2017.

[160] Q. Liu, Z. Xu, Y. Jiao, and M. Niethammer, "isegformer: Interactive segmentation via transformers with application to 3d knee mr images," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 464–474, Springer, 2022.

[161] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," arXiv preprint arXiv:2209.15076, 2022.

[162] X. Yu, Q. Yang, Y. Zhou, L. Y. Cai, R. Gao, H. H. Lee, T. Li, S. Bao, Z. Xu, T. A. Lasko, et al., "Unest: Local spatial representation learning with hierarchical transformer for efficient medical segmentation," arXiv preprint arXiv:2209.14378, 2022.

[163] Z. Xing, L. Yu, L. Wan, T. Han, and L. Zhu, "Nestedformer: Nested modality-aware transformer for brain tumor segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 140–150, Springer, 2022.

[164] Y. Tang, N. Zhang, Y. Wang, S. He, M. Han, J. Xiao, and R.-S. Lin, "Accurate and robust lesion recist diameter prediction and segmentation with transformers," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 535–544, Springer, 2022.

[165] Y. Li, S. Wang, J. Wang, G. Zeng, W. Liu, Q. Zhang, Q. Jin, and Y. Wang, "Gt u-net: A u-net like group transformer network for tooth root segmentation," in

International Workshop on Machine Learning in Medical Imaging, pp. 386–395, Springer, 2021.

[166] E. Sanderson and B. J. Matuszewski, "Fcn-transformer feature fusion for polyp segmentation," in Annual Conference on Medical Image Understanding and Analysis, pp. 892–907, Springer, 2022.

[167] Z. Zhao, Y. Jin, and P.-A. Heng, "Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery," in 2022 International Conference on Robotics and Automation (ICRA), pp. 11186–11193, IEEE, 2022.

[168] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," arXiv preprint arXiv:1902.03368, 2019.

[169] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations," in International conference on medical image computing and computer-assisted intervention, pp. 363–373, Springer, 2020.

[170] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, et al., "Nucleus segmentation across imaging experiments: the 2018 data science bowl," Nature methods, vol. 16, no. 12, pp. 1247–1253, 2019.

[171] T. S. Mathai, S. Lee, D. C. Elton, T. C. Shen, Y. Peng, Z. Lu, and R. M. Summers, "Lymph node detection in t2 MRI with transformers," in Medical Imaging 2022: Computer-Aided Diagnosis (K. M. Iftekharuddin, K. Drukker, M. A. Mazurowski, H. Lu, C. Muramatsu, and R. K. Samala, eds.), SPIE, apr 2022.

[172] Z. Shen, R. Fu, C. Lin, and S. Zheng, "COTR: Convolution in transformer network for end to end polyp detection," in 2021 7th International Conference on Computer and Communications (ICCC), IEEE, dec 2021.

[173] H. Li, L. Chen, H. Han, and S. K. Zhou, "SATr: Slice attention with transformer for universal lesion detection," in Lecture Notes in Computer Science, pp. 163–174, Springer Nature Switzerland, 2022.

[174] C. Niu and G. Wang, "Unsupervised contrastive learning based transformer for lung nodule detection," arXiv preprint arXiv:2205.00122, vol. 67, p. 204001, oct 2022.

[175] F. Shang, S. Wang, and Y. Yang, "An effective transformer-based solution for rsna intracranial hemorrhage detection competition," arXiv preprint arXiv:2205.07556, 2022.

[176] M. Zhou and S. Mo, "Shoulder implant x-ray manufacturer classification: exploring with vision transformer," arXiv preprint arXiv:2104.07667, 2021.

[177] H. Chen, C. Li, G. Wang, X. Li, M. M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, et al., "Gashis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection," Pattern Recognition, vol. 130, p. 108827, 2022.

[178] W. Liu, C. Li, M. M. Rahaman, T. Jiang, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, Y. Yao, et al., "Is the aspect ratio of cells important in deep learning? a robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers," Computers in biology and medicine, vol. 141, p. 105026, 2022.

[179] Q. Lyu, S. V. Namjoshi, E. McTyre, U. Topaloglu, R. Barcus, M. D. Chan, C. K. Cramer, W. Debinski, M. N. Gurcan, G. J. Lesser, et al., "A transformer-based deep learning approach for classifying brain metastases into primary organ sites using clinical whole brain mri images," arXiv preprint arXiv:2110.03588, 2021.

[180] F. Bertolini, A. Spallanzani, A. Fontana, R. Depenni, and G. Luppi, "Brain metastases: an overview," CNS oncology, vol. 4, no. 1, pp. 37–46, 2015.

[181] T. Stegmüller, A. Spahr, B. Bozorgtabar, and J.-P. Thiran, "Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification," arXiv preprint arXiv:2202.07570, 2022.

[182] M. Bhattacharya, S. Jain, and P. Prasanna, "Radiotransformer: A cascaded global-focal transformer for visual attention-guided disease classification," arXiv preprint arXiv:2202.11781, 2022.

[183] F. Zhang, T. Xue, W. Cai, Y. Rathi, C.-F. Westin, and L. J. O'Donnell, "Tractoformer: A novel fiber-level whole brain tractography analysis framework using spectral embedding and vision transformers," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 196–206, Springer, 2022.

[184] J. Zhang, Y. Nie, J. Chang, and J. J. Zhang, "Sig-former: monocular surgical instruction generation with transformers," International Journal of Computer Assisted Radiology and Surgery, pp. 1–8, 2022.

[185] J. Pang, C. Jiang, Y. Chen, J. Chang, M. Feng, R. Wang, and J. Yao, "3d shuffle-mixer: An efficient context-aware vision learner of transformer-mlp paradigm for dense prediction in medical volume," arXiv preprint arXiv:2204.06779, 2022.

[186] D. Reisenbüchler, S. J. Wagner, M. Boxberg, and T. Peng, "Local attention graph-based transformer for multi-target genetic alteration prediction," arXiv preprint arXiv:2205.06672, 2022.

[187] S. Płotka, M. K. Grzeszczyk, R. Brawura-Biskupski-Samaha, P. Gutaj, M. Lipa, T. Trzciński, and A. Sitek, "Babynet: Residual transformer module for birth weight prediction on fetal ultrasound video," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 350–359, Springer, 2022.

[188] H. H. Nguyen, S. Saarakkala, M. B. Blaschko, and A. Tiulpin, "Climat: Clinically-inspired multi-agent transformers for disease trajectory forecasting from multi-modal data," arXiv preprint arXiv:2104.03642, 2021.

[189] Y. Xie and Q. Li, "A review of deep learning methods for compressed sensing image reconstruction and its medical applications," Electronics, vol. 11, no. 4, p. 586, 2022.

[190] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Cukur, "Unsupervised mri reconstruction via zero-shot learned adversarial transformers," IEEE Transactions on Medical Imaging, 2022.

[191] W. Huang, P. Hand, R. Heckel, and V. Voroninski, "A provably convergent scheme for compressive sensing under random generative priors," Journal of Fourier Analysis and Applications, vol. 27, no. 2, pp. 1–34, 2021.

[192] J. P. Haldar and J. Zhuo, "P-loraks: low-rank modeling of local k-space neighborhoods with parallel imaging data," Magnetic resonance in medicine, vol. 75, no. 4, pp. 1499–1514, 2016.

[193] J. P. Haldar, "Loraks: Low-rank modeling of local k-space neighborhoods," in Proc. Int. Soc. Magn. Reson. Med, 2014.

[194] S. U. Dar, M. Yurt, M. Shahdloo, M. E. Ildız, B. Tınaz, and T. Çukur, "Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 6, pp. 1072–1087, 2020.

[195] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," Magnetic resonance in medicine, vol. 84, no. 6, pp. 3172–3191, 2020.

[196] D. Narnhofer, K. Hammernik, F. Knoll, and T. Pock, "Inverse gans for accelerated mri reconstruction," in Wavelets and Sparsity XVIII, vol. 11138, pp. 381–392, SPIE, 2019.

[197] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110–8119, 2020.

[198] C.-M. Feng, Y. Yan, H. Fu, L. Chen, and Y. Xu, "Task transformer network for joint mri reconstruction and super-resolution," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 307–317, Springer, 2021.

[199] P. Guo, Y. Mei, J. Zhou, S. Jiang, and V. M. Patel, "Reconformer: Accelerated mri reconstruction using recurrent transformer," arXiv preprint arXiv:2201.09376, 2022.

[200] J. Huang, Y. Wu, H. Wu, and G. Yang, "Fast mri reconstruction: How powerful transformers are?," arXiv preprint arXiv:2201.09400, 2022.

[201] Y. Long, Z. Li, C. H. Yee, C. F. Ng, R. H. Taylor, M. Unberath, and Q. Dou, "E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 415–425, Springer, 2021.

[202] C. Wang, K. Shang, H. Zhang, Q. Li, Y. Hui, and S. K. Zhou, "Dudotrans: Dual-domain transformer provides more attention for sinogram restoration in sparse-view ct reconstruction," arXiv preprint arXiv:2111.10790, 2021.

[203] J. Pan, H. Zhang, W. Wu, Z. Gao, and W. Wu, "Multi-domain integrative swin transformer network for sparse-view tomographic reconstruction," Patterns, p. 100498, 2022.

[204] T. Razi, M. Niknami, and F. A. Ghazani, "Relationship between hounsfield unit in ct scan and gray scale in cbct," Journal of dental research, dental clinics, dental prospects, vol. 8, no. 2, p. 107, 2014.

[205] S. N. Duda, N. Kennedy, D. Conway, A. C. Cheng, V. Nguyen, T. Zayas-Cabán, and P. A. Harris, "HL7 FHIR-based tools and initiatives to support clinical research: a scoping review," Journal of the American Medical Informatics Association, vol. 29, pp. 1642–1653, jul 2022.

[206] F. Auer, Z. Abdykalykova, D. Müller, and F. Kramer, "Adaptation of HL7 FHIR for the exchange of patients' gene expression profiles," in Studies in Health Technology and Informatics, IOS Press, jun 2022.

[207] C. Carter, B. Veale, et al., Digital radiography and PACS E-Book. Elsevier Health Sciences, 2022.

[208] M. D. Twa and C. A. Johnson, "Digital imaging and communication standards," Optometry and Vision Science, vol. 99, pp. 423–423, may 2022.

[209] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in Machine Learning in Medical Imaging, pp. 673–680, Springer International Publishing, 2019.

[210] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021.

[211] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, jul 2017.

[212] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pp. 72–82, Springer International Publishing, 2021.

[213] M. Xu, M. Islam, C. M. Lim, and H. Ren, "Learning domain adaptation with model calibration for surgical report generation in robotic surgery," in 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, may 2021.

[214] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287–1289, 2019.

[215] K. Papangelou, K. Sechidis, J. Weatherall, and G. Brown, "Toward an understanding of adversarial examples in clinical trials," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 35–51, Springer, 2018.

[216] P. Benz, S. Ham, C. Zhang, A. Karjauv, and I. S. Kweon, "Adversarial robustness comparison of vision transformer and mlp-mixer to cnns," arXiv preprint arXiv:2110.02797, 2021.

[217] T. Chuman and H. Kiya, "Security evaluation of block-based image encryption for vision transformer against jigsaw puzzle solver attack," in 2022 IEEE

4th Global Conference on Life Sciences and Technologies (LifeTech), pp. 448–451, IEEE, 2022.

[218] M. Li, D. Han, D. Li, H. Liu, and C.-C. Chang, "Mfvt: an anomaly traffic detection method merging feature fusion network and vision transformer architecture," EURASIP Journal on Wireless Communications and Networking, vol. 2022, no. 1, pp. 1–22, 2022.

[219] C. M. K. Ho, K.-C. Yow, Z. Zhu, and S. Aravamuthan, "Network intrusion detection via flow-to-image conversion and vision transformer classification," IEEE Access, vol. 10, pp. 97780–97793, 2022.

[220] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8, IEEE, 2021.

[221] K. D. Doan, Y. Lao, P. Yang, and P. Li, "Defending backdoor attacks on vision transformer via patch processing," arXiv preprint arXiv:2206.12381, 2022.

[222] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," Advances in Neural Information Processing Systems, vol. 34, pp. 8583–8595, 2021.

[223] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," arXiv preprint arXiv:2104.10972, 2021.

[224] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009.

[225] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.

[226] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform resnets without pre-training or strong data augmentations," arXiv preprint arXiv:2106.01548, 2021.

[227] H. Gani, M. Naseer, and M. Yaqub, "How to train vision transformer on small-scale datasets?," arXiv preprint arXiv:2210.07240, 2022.

[228] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," Medical Image Analysis, vol. 81, p. 102559, oct 2022.

[229] J. Lu, X. S. Zhang, T. Zhao, X. He, and J. Cheng, "April: Finding the achilles' heel on privacy for vision transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10051–10060, 2022.

[230] M. H. Mozaffari, S. Wen, N. Wang, and W.-S. Lee, "Real-time automatic tongue contour tracking in ultrasound video for guided pronunciation training.," in VISIGRAPP (1: GRAPP), pp. 302–309, 2019.

[231] M. Nandagopal, K. Seerangan, T. Govindaraju, N. E. Abi, B. Balusamy, and S. Selvarajan, "A deep auto-optimized collaborative learning (dacl) model for disease prognosis using ai-iomt systems," Scientific Reports, vol. 14, no. 1, p. 10280, 2024.

[232] A. Kernberg, J. A. Gold, and V. Mohan, "Using chatgpt-4 to create structured medical notes from audio recordings of physician-patient encounters: Comparative study," Journal of Medical Internet Research, vol. 26, p. e54419, 2024.

[233] F. Seyghalani Talab, B. Ahadinezhad, O. Khosravizadeh, and M. Amerzadeh, "A model of the organizational resilience of hospitals in emergencies and disasters," BMC Emergency Medicine, vol. 24, no. 1, pp. 1–13, 2024.

[234] HL7FHIR, "6.1.0 fhir security. available at https://build.fhir.org/security.html. [accessed online: Feb 16, 2023]."

[235] S. Zhang, S. Yang, G. Zhu, E. Luo, J. Zhang, and D. Xiang, "A fine-grained access control scheme for electronic health records based on roles and attributes," in Ubiquitous Security: First International Conference, UbiSec 2021, Guangzhou, China, December 28–31, 2021, Revised Selected Papers, pp. 25–37, Springer, 2022.

[236] M. Rashid, S. A. Parah, A. R. Wani, and S. K. Gupta, "Securing e-health iot data on cloud systems using novel extended role-based access control model," Internet of Things (IoT) Concepts and Applications, pp. 473–489, 2020.

[237] S. Khan, W. Iqbal, A. Waheed, G. Mehmood, S. Khan, M. Zareei, and R. R. Biswal, "An efficient and secure revocation-enabled attribute-based access control for ehealth in smart society," Sensors, vol. 22, no. 1, p. 336, 2022.

[238] M. W. Sanders and C. Yue, "Mining least privilege attribute based access control policies," in Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC '19, (New York, NY, USA), p. 404–416, Association for Computing Machinery, 2019.

[239] M. N. Nobi, R. Krishnan, Y. Huang, and R. Sandhu, "Administration of machine learning based access control," in Computer Security–ESORICS 2022: 27th European Symposium on Research in Computer Security, Copenhagen, Denmark, September 26–30, 2022, Proceedings, Part II, pp. 189–210, Springer, 2022.

[240] M. N. Nobi, R. Krishnan, Y. Huang, M. Shakarami, and R. Sandhu, "Toward deep learning based access control," in Proceedings of the Twelveth ACM Conference on Data and Application Security and Privacy, ACM, apr 2022.

[241] Z. Jin, L. Xing, Y. Fang, Y. Jia, B. Yuan, and Q. Liu, "P-verifier," in Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, ACM, nov 2022.

[242] A. Outchakoucht, E.-S. Hamza, and J. P. Leroy, "Dynamic access control policy based on blockchain and machine learning for the internet of things," International Journal of Advanced Computer Science and Applications, vol. 8, no. 7, 2017.

[243] A. Chiquito, U. Bodin, and O. Schelén, "Attribute-based approaches for secure data sharing in industrial contexts," IEEE Access, vol. 11, pp. 10180–10195, 2023.

[244] V. C. Hu, D. R. Kuhn, D. F. Ferraiolo, and J. Voas, "Attribute-based access control," Computer, vol. 48, no. 2, pp. 85–88, 2015.

[245] A. Ghorbani, A. H. LASHKARI, M. S. I. Mamun, and G. D. Gil, "Systems and methods for cybersecurity risk assessment of users of a computer network," July 30 2020. US Patent App. 16/753,301.

[246] Y. Wu, L. Li, B. Xin, Q. Hu, X. Dong, and Z. Li, "Application of machine learning in personalized medicine," Intelligent Pharmacy, vol. 1, no. 3, pp. 152–156, 2023.

[247] D. Guo, Applying Medical Language Models to Medical Image Analysis. PhD thesis, UCLA, 2024.

[248] Z. Lu, "Multimodal large language models in vision and ophthalmology," Investigative Ophthalmology & Visual Science, vol. 65, no. 7, pp. 3876–3876, 2024.

[249] J. Shapiro, S. Baum, F. Pavlotzky, Y. B. Mordechai, A. Barzilai, T. Freud, and R. Gershon, "Application of an nlp ai tool in psoriasis: A cross-sectional comparative study on identifying affected areas in patients' data," Clinics in Dermatology, 2024.

[250] D. He, T. J. Prabhakaran, E. Wang, and S. T. Chung, "Analyzing electronic medical records of low vision patients using a natural language processing framework," Investigative Ophthalmology & Visual Science, vol. 65, no. 7, pp. 5472–5472, 2024.

[251] I. C. Wiest, M.-E. Lessmann, F. Wolf, D. Ferber, M. Van Treeck, J. Zhu, M. P. Ebert, C. B. Westphalen, M. Wermke, and J. N. Kather, "Anonymizing medical documents with local, privacy preserving large language models: The llm-anonymizer," medRxiv, pp. 2024–06, 2024.

[252] M. A. Gismelbari, I. I. Vixnin, G. M. Kovalev, and E. E. Gogolev, "Speech emotion recognition using deep learning," in 2024 XXVII International Conference on Soft Computing and Measurements (SCM), pp. 380–384, IEEE, 2024.

[253] H. Jiang, "Research on emotion management for elderly based on speech signal analysis technology," in Fourth International Conference on Sensors and Information Technology (ICSI 2024), vol. 13107, pp. 1026–1033, SPIE, 2024.

[254] Y. Jin, M. Chandra, G. Verma, Y. Hu, M. De Choudhury, and S. Kumar, "Ask me in english instead: Cross-lingual evaluation of large language models for healthcare queries," in The Web Conference 2024.

[255] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, et al., "A large language model for electronic health records," NPJ digital medicine, vol. 5, no. 1, p. 194, 2022.

[256] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria, "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," arXiv preprint arXiv:2310.05694, 2023.

[257] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023.

[258] K. W. Church, "Word2vec," Natural Language Engineering, vol. 23, no. 1, pp. 155–162, 2017.

[259] P. T. Hung and K. Yamanishi, "Word2vec skip-gram dimensionality selection via sequential normalized maximum likelihood," Entropy, vol. 23, no. 8, p. 997, 2021.

[260] M. Habib, M. Faris, A. Alomari, and H. Faris, "Altibbivec: a word embedding model for medical and health applications in the arabic language," IEEE Access, vol. 9, pp. 133875–133888, 2021.

[261] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543, 2014.

[262] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

[263] U. D. HHS, "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule," 2023. Last accessed 05 Mar 2023.

[264] J. Walonoski, S. Klaus, E. Granger, D. Hall, A. Gregorowicz, G. Neyarapally, A. Watson, and J. Eastman, "Synthea™ novel coronavirus (covid-19) model and synthetic data set," Intelligence-based medicine, vol. 1, p. 100007, 2020.

[265] F. Gebali and M. Mamun, "Sram physically unclonable functions for smart home iot telehealth environments," 2022.

[266] L. Ménard, J. Aubin, M. Thibeault, and G. Richard, "Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model," Folia Phoniatrica et Logopaedica, vol. 64, no. 2, pp. 64–72, 2012.

[267] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," arXiv preprint arXiv:1811.12808, 2018.

[268] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[269] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (ssim) revisited: A data-driven approach," Expert Systems with Applications, vol. 189, p. 116087, 2022.

[270] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009.

[271] M. Bansal, M. Kumar, M. Sachdeva, and A. Mittal, "Transfer learning for image classification using vgg19: Caltech-101 image data set," Journal of Ambient Intelligence and Humanized Computing, pp. 1–12, 2021.

[272] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in International workshop on simulation and synthesis in medical imaging, pp. 1–11, Springer, 2018.

Appendix A

Appendix: Additional Information

A.1 Contrastive learning

1. Two data augmentation operations are applied to each of N randomly sampled minibatch samples, resulting in 2N augmented samples total.

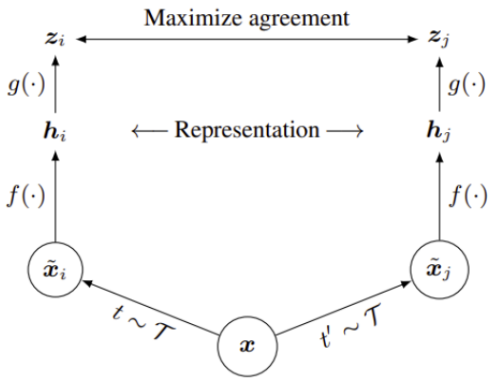
$$\tilde{x}_i = t(x), \qquad \tilde{x}_j = t'(x) \tag{A.1}$$

   where t,t' resembles the two data augmentation that are sampled from the same augmentation style τ. Various data augmentation techniques are used, such as random cropping, resizing with random flip, colour distortion, and Gaussian blur.

2. The representation that generated by encoder $f(.)$ assign two pairs as a positive samples. The reset $2(N-1)$ data points are assigned as negative samples.

$$h_j = f(\hat{x}_j), \qquad h_i = f(\hat{x}_i) \tag{A.2}$$

3. Cosine similarity $sim(x_i, x_j)$ used to derive the contrastive loss



(a)

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of $f, g, \mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^{\top} \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

(b)

Figure A.1: (a). The flow chart of contrastive learning process using *SimCLR* [9]. *SimCLR* maximizes agreement for the two input images x_i and x_j. While $f(.)$ is the encoder head, $g(.)$ is the projection head, and h is the feature representation of the image. (b). SimCLR Algorithm.

A.2 Data category example

Table A.1: Example of some attributes from the generated dataset categories.

Attribute example	Category
Position	User
Department	User
Speciality	User
Id	User
Insurance number	User
User access level	User
User consent	User
Password	User
Manager Id	User
MAC address	Device
IP address	Device
Device model	Device
Device type	Device
Device location	Device
Device manufacture	Device
Data category	Data
Data encryption	Data
Data storage location	Data
Storage type (Ex. Container, SSD, VM...)	Data
Data sensitivity level	Data
Data compliance	Data

A.3 Role-based access control system important factors

Table A.2: Examples of Role-based access control system data sources information.

Information source	Example of data source information factors
Client or operator	User id, role, department, level of access, geographic location
Patient	Patient ID, clinical condition, department, family doctor name, patient consent policy
Resources	Confidentiality, sensitivity, type of data, date ranges covered by the data, author of the data
Data context	System identity, transaction time, the expiration time of token data, the scope and purpose of the token, security of transaction