An innovative ecosystem based on deep learning : Contributions for the prevention and prediction of diabetes complications

Mohammud Shaad Ally Toofanee

Thèse de Doctorat

Université de Limoges

Université de Limoges
ED 653 – Sciences et Ingénierie

Thèse pour obtenir le grade de
Docteur de l'Université de Limoges
Informatique

Présentée et soutenue par
Mohammud Shaad Ally Toofanee
Le 20 Décembre 2023

An Innovative Ecosystem Based on Deep Learning: Contributions for the Prevention and Prediction of Diabetes Complications

Thèse dirigée par Damien Sauveron et Karim Tamine

JURY:

Président du jury
Samia Bouzefrane, Professeure des Universités, Laboratoire Cédric, CNAM, Paris

Rapporteurs
Marie Beurton-Aimar, Maître de Conférences, HDR, Laboratoire LaBRI, Université de Bordeaux
Faten Chaieb-Chakchouk, Professeure des Universités, Laboratoire Efrei, Ecole EFREI Paris

Examinateurs
Damien Sauveron, Professeur des Universités, XLIM, Université de Limoges
Karim Tamine, Maître de Conférences, XLIM, Université de Limoges

Invités
Xavier Debussche, Endocrinologue, Centre Hospitalier de Paimpol, Chercheur INSERM
Mohamed Hamroun,Mohamed Hamroun, Maître de Conférences, XLIM, 3IL

To the Toofanee, Dowlut and Kurreemun Families

We ought not to be embarrassed of appreciating the truth and of obtaining it wherever it comes from, even if it comes from races distant and nations different from us...

Al-Kindi

Acknowledgements

Une thèse n'est pas un exercice solitaire. Elle est le fruit de la collaboration et du soutien de nombreuses personnes et institutions qui la rendent possible. Je tiens à accorder toute l'importance nécessaire à cette section de remerciements afin de n'oublier personne.

Institutions

Je tiens à exprimer ma profonde gratitude envers la Présidente de l'Université des Mascareignes, Pr. Nathalie Bernardie-Tahir, qui a su créer les conditions nécessaires pour le démarrage de cette thèse. Je souhaite remercier la Présidente de l'Université de Limoges, Pr. Isabelle KLOCK-FONTANILLE, pour la mise en place du projet ULIMA, qui a rendu possible les échanges et les mobilités essentielles à cette recherche. Je souhaite adresser mes remerciements à l'administration de l'Université des Mascareignes : notamment la responsable des finances, le registrar, le doyen de la faculté TIC ainsi que le directeur général pour leur précieux soutien logistique.
Je remercie l'équipe de recherche EpiMaCT (Épidémiologie des maladies chroniques en zone tropicale) car j'ai eu accès à un serveur avec un GPU qui a permis de faire tourner des algorithmes qui ont parfois pris entre une semaine et un mois pour se terminer. Tous les algorithmes des publications ont été exécutés sur ce serveur.
Je remercie le personnel de l'école doctorale pour leur assistance dans les inscriptions, réinscriptions et l'orientation fournie pour naviguer sur ADUM, en particulier Sabrina BRUGIER. Mes remerciements vont également à la directrice de l'école doctorale, Pr. Anne JULIEN-VERGONJANNE, pour sa gentillesse et sa bienveillance durant le CSI. Merci à mes collègues d'XLIM qui m'ont chaleureusement accueilli durant mes immersions au laboratoire XLIM.
En tant que membre de l'Union des enseignants-chercheurs, les collègues de l'équipe dirigeante de l'Union, UDMASU, ont souvent excusé mes absences aux réunions. Je tiens à leur exprimer également toute ma gratitude.
Je n'oublie pas de remercier mes collègues du département, Khadim RAMOTH, Soonita SEEBURRUN, Abdel HOSSENDBUX, et la chef de département, Sabeena DOWLUT, pour les ajustements qu'ils ont acceptés afin de garantir le bon déroulement de cette thèse.

Encadrants

J'exprime ma sincère reconnaissance envers mes directeurs de thèse, le Pr. Damien SAUVERON, pour son précieux soutien tout au long de cette thèse, ainsi que le Dr. Karim TAMINE, qui

fut de tous les combats et surtout la confiance qu'il m'a témoignée dans les moments les plus compliqués sans jamais mettre un mot de travers. Je remercie aussi Mohamed HAMROUN pour son implication dans l'encadrement.

Amis

Cette thèse m'a permis d'élargir mon cercle d'amis, et je souhaite exprimer ma gratitude envers ceux qui ont joué un rôle essentiel. Je tiens à mentionner :

- Nathalie BERNARDIE-TAHIR, grâce à qui toute cette aventure a pu commencer.
- Ouiddad LABANNI-IGBIDA pour son encouragement et son accueil chaleureux au sein de sa famille, qui a rendu l'éloignement de Maurice moins difficile.
- Je tiens à exprimer ma gratitude envers Farid BOUMEDIENE pour la confiance qu'il m'a accordée, ses encouragements constants, et nos discussions continues sur la thèse, les publications et la recherche. Ses paroles résonnent encore dans ma tête : "La thèse, c'est trois ans !!!" et "Il y a une vie après la thèse".
- Vincent PETIT pour nos discussions en fin de journée après avoir travaillé sur des scripts VBA, ainsi que pour son soutien dans la mise en œuvre des idées novatrices et la charge de travail fourni par son serveur.
- Mohamed HAMROUN pour notre collaboration pour les publications et aussi les discussions concernant les projets futurs.
- Lynda TAMINE-LECHANI, pour avoir révisé nos articles depuis la première version jusqu'à la dernière, en fournissant des commentaires et des recommandations pertinentes.
- Kader et Dahbia ALIBENALI, pour leur soutien médical, leur gentillesse, leur accueil, et les discussions philosophiques.
- Mériem BENMEHEL pour son accueil chaleureux et sa gentillesse en me recevant dans son appartement.

La Famille

Malgré la présence de tous les éléments mentionnés précédemment, s'il y a un élément qui est véritablement la pièce maîtresse de cette thèse, c'est ma famille. Je tiens à rendre hommage à mon oncle, Pharad KURREEMUN, qui n'est malheureusement plus parmi nous, mais qui a toujours été une source d'inspiration et le restera à jamais. Mes pensées vont également à ma cousine, Parween TOOFANEE, qui nous a quittés pendant mon absence du pays. Je tiens à remercier chaleureusement Karim et Maya, qui sont devenus ma famille en France, ainsi que Madina et Redha, ma famille à Paris. Un grand merci à Salim et Mazan pour avoir toujours répondu présents quand nous avions besoin d'eux. Mes sœurs, Kawsar et Farzanah, ont été

omniprésentes pour prêter main-forte, que je sois au pays ou à l'étranger. Mes beaux-parents, Sabir, Shirin, et mon beau-frère, Ikbal Dowlut, méritent toute ma reconnaissance pour leur soutien et leur responsabilité envers mes enfants.
Je dédie toutes mes réussites à mon papa, Rashid TOOFANEE, laboureur, et à ma maman, Nazma KURREEMUN, ouvrière dans une usine textile.
Un hommage particulier à la chef de famille, véritable superwoman, Sabeena, mon épouse, qui est la pierre angulaire de cette thèse. Enfin, je tiens à exprimer ma profonde gratitude envers ceux qui ont fait les plus grands sacrifices : mes enfants, Zainab, 13 ans ; Saima, 11 ans ; Suffyaan, 10 ans ; et Zakia, 8 ans.

Droits d'auteurs

Table of Contents

Table of Figures

List of Tables

List of Abbreviations

4G – Fourth Generation Mobile Radio.

ADA – American Diabetes Association.
AI – Artificial Intelligence.
ANN – Artificial Neural Network.

BERT – Bidirectional Encoder Representation from Transformers.

CKD – Chronic Kidney Disease.
CNN – Convolutional Neural Network.

DFU – Diabetic Foot Ulcer.
DFUs – Diabetic Foot Ulcers.
DL – Deep Learning.
DNN – Deep Neural Network.
DNNs – Deep Neural Networks.
DR – Diabetic Retinopathy.

EHRs – Electronic Health Records.

FL – Federated Learning.

GAN – Generative Adversarial Network.
GANs – Generative Adversarial Networks.
GPT – Generative Pre-trained Transformer.

IDF – International Diabetic Federation.

KNN – K-Nearest Neighbor.

LN – Learning Nest.

LR – Linear/Logistic Regression.
LSTM – Long Short-Term Memory.

ML – Machine Learning.
MLP – Multi-Layer perceptron.

NLP – Natural Language Processing.
NNs – Neural Networks.

RF – Random Forest.
RNN – Recurrent Neural Network.

SNN – Siamese Neural Network.
SNNs – Siamese Neural Networks.
SVM – Support Vector Machine.

T2D – Type 2 Diabetes.

ViT – Vision Transformer.

WHO – World Health Organisation.

1 Introduction

Summary

1.1 Context and Motivation

Diabetes is termed as 'Modern Preventable Pandemic' and continues to escalate, with presently 10% of the world population having Type 2 Diabetes (T2D) [1]. Diabetes, a chronic metabolic disorder characterized by elevated blood glucose levels, has emerged as one of the most pressing global health challenges. In 2021, an estimated 537 million people were living with diabetes, a number projected to soar to 643 million by 2030 and a staggering 783 million by 2045 [2]. According to the latest report from the International Diabetes Federation, 22.6% of the Mauritian population had diabetes in 2021, and this figure is projected to rise to 26.6% in 2045 [2]. One of the main causes of death in Mauritius in 2020 is complications resulting from diabetes mellitus, which is 21% [3]. Public health services in Mauritius are provided free of charge, and the budget for the Ministry for the Financial Year 2021-2022 was around MUR 13.1 billion, representing around 7% of government expenditure [4] and representing an increase of approximately 91.8% from 2010 [5].
There are three types of diabetes:

- Type 1 diabetes: It is the major type of diabetes in childhood but can occur at any age. It cannot be prevented. People with type 1 diabetes require insulin to survive [2].
- Type 2 diabetes(T2D): It accounts for the vast majority (over 90%) of diabetes worldwide. Evidence exists that type 2 diabetes can be prevented or delayed [2]. This thesis tackles issues related to T2D.
- Gestational diabetes: It is a temporary condition that occurs during pregnancy and carries a long-term risk of T2D.

Pre-diabetes is a condition characterized by elevated blood glucose levels below the threshold for a diagnosis of diabetes but associated with a higher risk of developing diabetes [6] or increased glycated hemoglobin A1c (HbA1c) levels. It indicates a higher risk of developing type 2 diabetes and diabetes-related complications [2].
The associated complications with diabetes are an increased incidence of cardiovascular-related diseases, kidney problems, diabetic retinopathy [7], and the development of Diabetic Foot Ulcers (DFUs). The early detection of diabetes and the initiation of treatment are extremely important in the management of diabetes and the prevention of complications. The longer a person has diabetes but remains undiagnosed, the greater the risk of developing complications [2].

Chronic diseases are especially taxing on the healthcare system when they afflict a large portion of the population for long periods of time. This adversely affects the well-being of many and constitutes a large portion of the cost to the healthcare system. Given the number of diabetic patients and its anticipated growth, it has become increasingly unfeasible for conventional healthcare systems to economically sustain the personalized care required for effective chronic disease management. The socio-economic impacts are very high and require non-conventional interdisciplinary interventions.

Figure 1.1: Illustration of the Learning Nest: Patients attend therapeutic education sessions in small groups, which involve nutritionists, medical doctors, and healthcare educators. The aim is to empower the patient. Based on the results of the blood test obtained before attending the session, a total health score is calculated, and depending on the score obtained, the patient is encouraged to decide on remedial lifestyle actions based on his individual context. The complications linked to diabetes and associated complications are also discussed, with the focus being on creating a conducive environment for patients to voice their concerns.

This thesis provides a pioneering approach at the intersection of healthcare, Artificial Intelligence (AI) and diabetes prevention, education, and management. It bridges the gap between a traditional diabetes care ecosystem, as shown in Figure 1.1, the Learning Nest (LN) [8], and the application of AI as a transformative technology to propose an AI-powered ecosystem.

1.1.1 AI and Healthcare

Before moving further into this thesis, it is crucial to clarify some core concepts that will be addressed in our study. AI is a field of computer science and engineering that revolves around the research and development of computer programs capable of executing tasks falling within the purview of human intelligence [9], [10]. It is important to clarify that the terms machine learning (ML) and deep learning (DL), often mistakenly used interchangeably, are actually hierarchical. Machine learning (ML) is a subfield of AI, whereas DL is a subset of ML.

ML algorithms possess the remarkable capacity to iteratively learn from domain-specific training data without the need for explicit programming, thereby uncovering insights that would otherwise remain hidden [11]. Frequently used algorithms in ML encompass artificial neural networks (ANN), support vector machines (SVM), k-nearest neighbors (k-NN), decision trees (DT), and naïve Bayes (NB) [12].
Deep learning (DL), on the other hand, represents a more specialized domain within ML. DL applies complex architectures such as convolutional neural networks (CNNs) [13], recurrent neural networks (RNNs) [14], and transformers [15], enabling it to tackle intricate tasks associated with extensive, heterogeneous, and high-dimensional datasets [16]. It is pertinent to highlight that, in addressing specific challenges, ML algorithms are at times integrated with deep learning architectures. DL has notably attained state-of-the-art performance, particularly in image and textual data processing tasks.
Artificial intelligence (AI) is rapidly transforming the healthcare sector, offering unprecedented opportunities to address the alarming challenges of public health. AI-powered tools and technologies are poised to revolutionize disease prevention, diagnosis, treatment, and management, empowering healthcare providers and patients alike. This is possible due to the growing abundance of healthcare data and the rapid advancements in analytics techniques [17]. The use of artificial intelligence (AI) in medicine is beginning to alter current procedures in prevention, diagnosis, treatment, amelioration, cure of disease, and other physical and mental impairments [18]. One of the key appeals of AI lies in its ability to emulate a diverse array of human-like functions, its capacity to learn from past experiences, and its adaptability to new inputs and evolving contexts [18]. With appropriate design and implementation, AI can enhance patient care while reducing expenditure on healthcare [19]. It can empower patients and communities to assume control of their own health care and better understand their evolving needs [20]. Furthermore, it can identify new relationships between genetic codes or control surgery-assisting robots [21]. During the recent unprecedented COVID-19 pandemic, AI played a pivotal role across diverse applications, encompassing disease diagnosis, predictive modeling of transmission dynamics, outcome forecasting, and research on effective and secure medicines and vaccines [22], [23], [24]. Notably, the potential of AI was underscored by the emergency authorization granted to an untested AI algorithm by the US Food and Drug Administration [24]. However, such an expedited approach stands in contrast to our vision for the integration of AI within the realm of healthcare. We advocate for a rigorous, multidisciplinary evaluation process as a prerequisite for the deployment of any AI-driven solution in the healthcare domain.

1.1.2 AI and Diabetes

The Artificial Intelligence (AI) revolution, propelled by cost-effective computational resources and vast data availability, has significantly influenced the domain of diabetes prevention,

education and management. As depicted in Figure 1.2, Figure 1.2 shows recent and future medical innovations to help people living with diabetes [25].

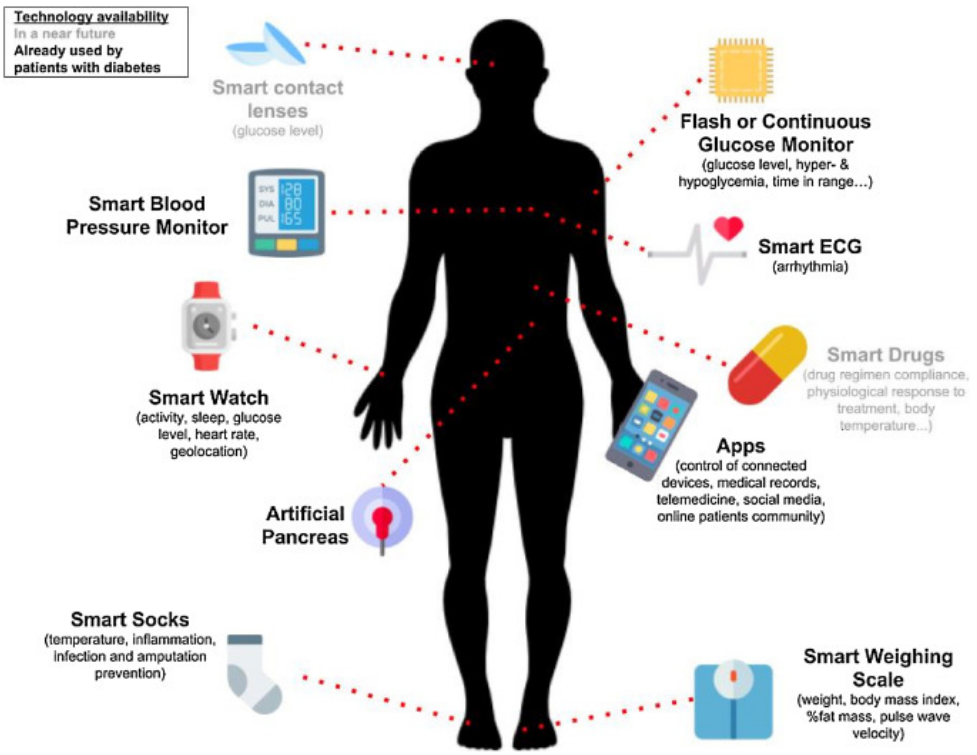


Figure 1.2: Recent and future medical innovations to help people living with diabetes. ECG: electrocardiography [25].

AI tools is used for Diabetes management, personalised nutrition for prevention and management [26], [27], as virtual assistant for doctors and social networks patient group interaction monitoring. Medical imaging and Deep Learning (DL) are being applied in the field of retinopathy [28], [29].
Several research, as evidenced by the systematic review [30], has delved into predicting cardiovascular diseases in diabetic patients, given their increased risk to cardiovascular complications due to their underlying condition. Concurrently, some studies have centered on prediction on-set of diabetes [31] by comparing various machine learning methodologies. In a recent study, Al-Tawil *et al.et al.* [32] employed bio-inspired metaheuristic algorithms for feature selection, targeting two unique datasets tailored for diabetes categorization. Similarly, Khaleel *et al.* [33] explored the PIMA dataset to prognosticate diabetes, employing Logistic Regression (LR), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) methodologies. Furthermore, a recent study [34] adopted a machine learning paradigm to categorize, detect early stages, and predict diabetes utilizing clinical data such as blood glucose levels, blood pressure, and body mass index. This research utilized three distinct machine learning classifiers: random forest (RF), multilayer perceptron (MLP), and logistic regression (LR), aiming to categorize diabetic from non-diabetic individuals.

1.1.3 Deep Learning and Diabetes

Deep learning (DL) algorithms have shown great promise compared to traditional ML algorithms. State-of-the-art outcomes have been realized using pre-trained architectures, particularly in the domains of imaging and textual data processing, and hence medical images and health-related massive text. In the context of diabetes management and research, DL has been applied to:

1. predict, detect, and classify complications linked to diabetes, which are cardiovascular events, diabetic retinopathy, and diabetic foot ulcers (DFUs).
2. predict the early onset of diabetes using biomedical data.
3. explore massive text data to educate patients and find insight for stakeholders dealing with the fight against diabetes.

DL can significantly improve healthcare outcomes by increasing the speed, accuracy, and cost-effectiveness of interpreting medical images [35]. Zhu *et al.* [36] reviewed 40 research studies and concluded that DL techniques outperform conventional ML approaches concerning many diabetes-related While a radiologist's knowledge is based on a limited number of cases they have encountered, DLs can learn from vast amounts of data to make diagnoses and predictions. Bandary *et al.* [37] review the detection of early-stage diabetic retinopathy (DR) using DL. DR can lead to permanent sight loss. Their review shows that DL techniques dominate as techniques applied for early detection of DR, and this is also confirmed by Sebastian *et al.* [38]. Furthermore, in a very recent study by Das *et al.* [39], it was again demonstrated that DL outperforms traditional ML models.

While the application of DL models to imaging has been highlighted, it is essential to recognize that DL techniques also play a pivotal role in Natural Language Processing (NLP) for textual data analysis. For instance, Ahne *et al.et al.* [40] harnessed DL, specifically the BERT-base models, to derive insights from diabetes-associated tweets spanning 2017 to 2021. In another exploration, Anoop *et al.* [41] employed transformers for sentiment analysis derived from tweets. Adding to this, Vidyadharan *et al.* [42] integrated DL with NLP to undertake an evidence-based research approach towards diabetes prevention and management. Similarly, Yu *et al.* [43] implemented advanced models like Bidirectional Encoder Representation from Transformers (BERT) [44], Robustly Optimized BERT Pretraining Approach (RoBERTa) [45], and a Recurrent Neural Network (RNN) [14] to pinpoint relevant concepts within clinical narratives.

1.2 Research Questions

In this thesis we work toward the integration of Deep Learning techniques with the ecosystem described by the Learning Nest [46] and we consequently address three main research questions (RQs) in this thesis:

RQ-1: How can deep learning be applied for DFU management?

RQ-2: How to ensure the confidentiality of healthcare data used for training deep learning models?

RQ-3: How to use deep learning models to implement chatbot capable of addressing inquiries related to diabetes?

1.3 Thesis Contribution and Findings

DFU is one of the most significant and devastating complications of diabetes which aggravates the patient's condition while also having a significant socioeconomic impact [47]. Ulceration of the foot in diabetes is common and disabling and frequently leads to amputation of the leg, and the mortality is high [48]. When infection complicates a foot ulcer, the combination can be limb- or life-threatening [49]. This is the need to detect at the earliest possible time, and once treatment starts, do a close follow-up. Early detection and close monitoring of treatment are imperative in addressing this issue. While previous research in this domain has predominantly proposed Convolutional Neural Network (CNN) based approaches, our approach combines CNN and Vision Transformers (VIT) within Siamese Neural Network (SNN) for multiclass classification of DFU into four distinct classes: None, Infection, Ischaemia, and Both. This combination yielded superior results compared to other studies in this field. We further exploited this architecture, and we propose an innovative system capable of effectively facilitating the longitudinal follow-up of DFU patients after the start of treatment. This system exhibits the potential to assist healthcare professionals and, in the future, be extended to mobile devices for direct patient engagement in preventive measures. We also address challenges related to interpretability, ethics, and the pressing need for improved dataset quality. One proposed solution is the collection of data from multiple healthcare facilities to enhance dataset quality and research outcomes.
After achieving a classification model with satisfactory performance, our research shifted focus towards positively influencing the LN ecosystem and aiding healthcare professionals. Consequently, we suggest leveraging the similarity learning capabilities of SNN to offer a tool tailored for the continual monitoring of treatment protocols. Our findings confirm the system's efficacy, emphasizing that enhanced data quality can further improve its performance.
The use of AI in healthcare poses significant challenges related to ethics and privacy [50]. One of the major challenges is the collection of healthcare data in one country while storing it on remote servers in another country with different legal jurisdictions [51]. Another challenge is the effective governance of such data, ensuring careful aggregation and appropriate access to drive

innovation, enhance patient outcomes, and improve the efficiency of healthcare systems, all while safeguarding the privacy and security of individuals' data [52]. State-of-the-art algorithms often necessitate access to extensive datasets, and these datasets may not always be physically located in the same location or controlled by a single entity. This situation raises significant concerns regarding the confidentiality of the data, and it is imperative to address these concerns during the design and training phases of AI systems. We thus propose the use of centralized federated learning (FL) and also peer-to-peer FL, which we test on our DFU dataset using three distinct heuristics for P2P FL. The preliminary results tend to show that performance in terms of accuracy is comparable for P2P FL.
Diabetes education plays a crucial role in empowering patients to manage their condition effectively, improving both their quality of life and health outcomes. Several studies have highlighted the benefits of utilizing conversational agents in various healthcare environments. These benefits include promoting behavioral modifications and guiding individuals towards healthier living habits [53], [54], [55]. Hence, we propose an innovative chatbot pipeline for keyword generation, question generation, question answering, and a chatbot fine-tune on a large language model based on BERT. We also propose a pipe and algorithm to create a dataset based on context, questions, and answers on diabetes. We also create a dataset of context on diabetes. We used part of the dataset to conduct a survey on diabetes in online media. If the pipeline is fed local language text, we hope to capture the specificities of the local language in context. It should be noted that the chatbot will not do a clinical diagnosis. The scientific contributions proposed in this thesis serve to enrich the ecosystem of the Learning Nest [46] for the monitoring of diabetes patients in three different aspects: A tool for the Classification and tracking of the progression of a disease related to diabetes, taking into account the confidential aspect of patients' health data in the implementation of these monitoring tools, and finally, ongoing interaction with the patient to assist in managing their condition. The generic aspect of the deep learning architectures proposed in these contributions will allow for potential developments and extensions concerning not only diabetes but also other chronic diseases.

1.4 Thesis Layout

The remainder of this thesis consists of the following chapters:

Chapter 2 Literature Review

In chapter 2, we provide a comprehensive overview of the key AI-related technologies that were studied and some employed throughout the course of this thesis. An important relates to the internal working of Artificial Neural Network (ANN). We also explore DL models for

image classification and transformer-based models for NLP. We put effort into explaining the metrics that are applied for interpreting the performance of models.

Chapter 3 Diabetic Foot Ulcer and Machine Learning

In this chapter 3, we address RQ-1: "How can deep learning be effectively utilized for DFU management?" The content and findings presented herein are drawn from our collaborative publications, specifically [56] and [57].
We subdivide this chapter into two main sections:

1. Section 3.2 deals with proposing a novel architecture for DFU classification after analyzing the latest literature in this field.
2. Section 3.3 deals with exploiting the findings of the first part to propose the use of DL for the implementation of a tool that will help healthcare professionals with longitudinal follow-up of the DFU.

This chapter motivates two of our main contributions which are the DFU classifier and the DFU tool for helping healthcare professionals for treatment follow-up.

Chapter 4 Confidentiality of Healthcare Data

In this chapter 4, we tackle Research Question 2 (RQ-2): "How to ensure the confidentiality of healthcare data used for training deep learning models?". The issue of ensuring data confidentiality becomes particularly challenging when training deep learning models with data that is geographically dispersed. To ground the discussion, we first provide foundational knowledge on Federated Learning (FL) and Peer-to-Peer FL (P2P FL). Following this, we explore existing literature pertinent to the field. Our proposed solutions are then presented, accompanied by a detailed explanation of each heuristic set to be implemented. The chapter concludes with a presentation of the experiments carried out, the results achieved, an encompassing discussion, limitations observed, and potential directions for future research.

Chapter 5 AI Chatbot for Diabetes

Chapter 5 focuses on Research Question 3 (RQ-3): "How to use deep learning models to implement a chatbot capable of addressing inquiries related to diabetes?". We start off by explaining the motivating factors behind this research. This is followed by an introductory explanation on Natural Language Processing (NLP) to provide essential background. Subsequently, we present a review of pertinent literature in the domain. With that foundation, we present our proposed solution in detail. After laying out our approach, we highlight the experiments undertaken, showcasing the quantitative outcomes derived. The chapter ends in a comprehensive discussion, outlining the limitations of our study, and concludes by suggesting potential trajectories for future research to build upon this work.

Chapter 6 Conclusion and Perspectives

In this chapter, we consolidate and recapitulate the primary research contributions in alignment with the specified Research Questions (RQs), offering a holistic overview. Finally, we propose potential avenues and recommendations for future explorations in the field of AI and Diabetes.

Appendices

This thesis includes two appendices.

1. The first appendix is a publication related to a survey study we carried out: "Analyzing Online Media Articles on Diabetes using Natural Language Processing: A Comparative Study of Indian Ocean Region and France which compared how diabetes was dealt with in Indian ocean online media articles and France by applying NLP techniques". This is also in line with the LN and AI powered ecosystem.
2. The second appendix is a publication related to a study we conducted on using computer vision and deep learning for Mauritian Food Image classification: "DLMDish: Using Applied Deep Learning and Computer 6 Vision to Automatically Classify Mauritian Dishes".

1.5 Publications

1.5.1 Journals

1. **Toofanee, Mohammud Shaad Ally**, Sabeena Dowlut, Mohamed Hamroun, Karim Tamine, Vincent Petit, Anh Kiet Duong, and Damien Sauveron. "DFU-SIAM a Novel Diabetic Foot Ulcer Classification with Deep Learning." IEEE Access (2023).
2. **Toofanee, Mohammud Shaad Ally**, Sabeena Dowlut, Mohamed Hamroun, Karim Tamine, Anh Kiet Duong, Vincent Petit, and Damien Sauveron. "DFU-Helper: An Innovative Framework for Longitudinal Diabetic Foot Ulcer Diseases Evaluation Using Deep Learning." Applied Sciences 13, no. 18 (2023): 10310.
3. **Toofanee, Mohammud Shaad Ally**, Omar Boudraa, Tamine Karim. "DLMDISH: Using applied deep learning and computer vision to automatically classify mauritian dish." International Journal of Image and Graphics.
4. **Toofanee, Mohammud Shaad Ally**, Sabeena Dowlut, Mohamed Hamroun, Karim Tamine, Anh Kiet Duong, Vincent Petit, and Damien Sauveron. "Federated learning: Centralised and P2P to a Siamese deep learning model for Diabetes Foot Ulcer classification." Applied Science [Submitted]

1.5.2 Conferences

1. **Toofanee, Mohammud Shaad Ally**, Nabeelah Zainab Ally Pooloo, Sabeena Dowlut, Karim Tamine, and Damien Sauveron. "Analyzing Online Media Articles on Diabetes using Natural Language Processing: A Comparative Study of Indian Ocean Region and France." In CS & IT Conference Proceedings, vol. 13, no. 8. CS & IT Conference Proceedings, 2023.
2. **Toofanee, M. Shaad Ally**, B. Sabeena Dowlut, Maryvette Balcou-Debussche, Xavier Debussche, Veronique Lahausse, and Luqman Nisa. "A Mobile Application to Empower Diabetic Patients Enrolled in a Therapeutic Patient Education Programme in Mauritius." In 2022 IST-Africa Conference (IST-Africa), pp. 1-6. IEEE, 2022.

1.5.3 Oral and Poster Communications

1. **Toofanee Shaad**, Tamine K., "Artificial Intelligence for Diabetes Management and Education [AID-ME]", IA pour les Sciences de l'Ingénierie, Workshop Organised by l'Institut des sciences de l'ingénierie et des systèmes du CNRS, June 2022.
2. Oral Presentation: **Toofanee Shaad**, Tamine Karim, Dowlut Sabeena, Boumediene Farid, "Programme Smart DIABETE et Projet IoT HD". Congrès HopiPharm, Mai 2023, Strasbourg, France
3. Oral Presentation: Dowlut Sabeena, **Toofanee Shaad**, "Smart Diabetes Education", Ile Maurice. In 3ème Congrès de Recherche en Santé de l'Océan Indien. 2022.
4. Poster Presentation: Sabeena, Dowlut, **Shaad Toofanee**, Pierre-Marie Preux, Julien Magne, Clémence Thebaut, Abel Bellati, Yaasir Ozeer *et al.*et al. "Une intervention multi-composante évolutive mobilisant des objets connectés pour la gestion de l'hypertension et/ou diabète dans le contexte mauricien." In 3ème Congrès de Recherche en Santé de l'Océan Indien. 2022.
5. Oral Presentation: **Toofanee Mohammud Shaad Ally**, "Artificial Intelligence for Diabetes Management and Education". Fédération MIRES, Mathématiques et leurs Interactions, Images et Information numérique, Réseaux et Sécurité, Séminaire Axe 4-Sciences des Donneées. Juin 2022.
6. Poster Presentation: **Toofanee Shaad**, Tamine Karim, "Artificial Intelligence for Diabetes Management and Education [AID-ME]", Journée Interdisciplinarité, Université de Limoges.

2

Literature review

Summary

2.1 Introduction

In this particular section, we will be discussing the key concepts which are relevant and have been considered for this thesis. We will present and overall view of techniques which were considered to carry our the the various experimentation which includes Natural Language Processing, Machine Learning methods and metrics.

2.2 AI and Healthcare

An ageing populations, growing burden of chronic diseases and rising costs of healthcare around the world challenges governments, regulators and healthcare institutions to innovate and transform models of healthcare delivery [58]. AI has the potential to answer some of the problems mentioned but should respect some key elements. Key challenges are identified as data quality and access, technical infrastructure, organisational capacity, and ethical and responsible practices in addition to aspects related to safety and regulation [58].

2.2.1 Data for Healthcare and Privacy

The availability of reliable, massive data is important to harvest the benefits and promise of AI-associated technologies like machine learning and deep learning, known for thriving on big data. Medical data have the characteristics of disease diversity, heterogeneity of treatment and outcome, and the complexity of collecting, processing, and interpreting data [59]. Data from healthcare can come from the following sources [60]:

1. web and social media data.
2. reading from medical internet of things(MIoT) devices.
3. health care claims and other billing records.
4. biomedical data: genetics, retinal scans, x-ray and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data.
5. Human generated data: unstructured and semi-structured data such as EHRs, physicians notes, email, and paper documents.

Electronic Health Records (EHRs) systems store data associated with each patient encounter, including demographic information, diagnoses, laboratory tests and results, prescriptions, radiological images, and clinical notes [61]. These data are the raw materials used by researchers in AI and healthcare. The two main types of data that are manipulated are images and text data. For images we would like to include self-taken pictures via smart phones.
To ensure AI has the required impact and is able to positively revolutionize the healthcare system, fundamental issues have to be solved in terms of [62]:

- Who owns health data?
- Who is responsible for it?
- Who can use it?

AI algorithms require access to vast amounts of patient data, including sensitive health information. Protecting patient privacy and ensuring the secure storage and transmission of data are essential to maintaining patient trust and complying with regulatory requirements. When dealing with medical data, researchers usually choose to anonymize the data by coding the name and surname of the patient. This is not deemed acceptable for ensuring anonymity. Efforts should be put into the development of new anonymization processes where an individual can contribute to the development of models and adoption of patient care while also ensuring his or her rights to privacy [63]. A potential avenue is Federated Learning, which ensures data remains on consumer devices instead of moving to a remote, centralized server.

2.2.2 Ethical AI in Healthcare

ML and DL open possibilities that could emulate, and even surpass, human capabilities on certain tasks. These models need massive amounts of data for training. However, there are many groups or sub-groups that are not electronically present in any biomedical database and hence absent from the training data of these models, which increases real-life bias [64]. An example that shed light on the pivotal role of ethics is the use of AI for the detection of melanoma, where AI algorithms, having been trained predominantly on images of lighter skin shade making it less accurate for detection on dark-skinned people [65]. Hence, while acknowledging the potential of AI to enhance diagnosis, treatment, health research, and drug development, as well as support governments in carrying out public health functions such as surveillance and outbreak response, the WHO issued strict guidelines that need to be followed [66] among which ethical guidelines are as follows:

- Avoid harming others;
- Promote the well-being of others;
- Ensure that all persons are treated fairly;
- Deal with people in ways that respect their interests.

It is noteworthy that numerous public and private institutions across various countries have promulgated their own guidelines for employing AI. Nevertheless, due to the expansive nature of these guidelines and the utilization of ambiguous terms, their impact can be markedly limited, and some critics have provocatively deemed them futile and unattainable [67]. Murphy *et al.* [64] point to the complexity of ethical issues in the health sector and advise cautious optimism. In our opinion, the phenomenal speed at which AI technologies are advancing poses significant challenges for the development of policies and ensuring their implementation. Yet, it is heartening to observe that this issue is prominently discussed whenever AI is considered, indicating a growing awareness and emphasis on ethical considerations.

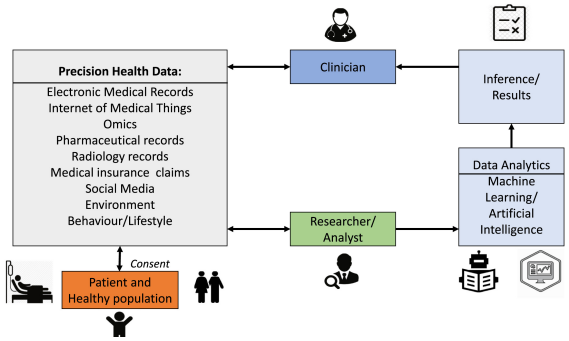


Figure 2.1: Precision health ecosystem from [68].

2.2.3 Explainable AI (XAI)

The majority of DL models are perceived as "black boxes" [69]. Within the healthcare domain, the interpretability of results is crucial; otherwise, the technology may encounter resistance in terms of acceptability. If a model provides a diagnosis that diverges from a radiologist's assessment, it is imperative that the model can explain its reasoning. For example, features used by a model to output a decision should be understandable to healthcare professionals. It should be noted that trustworthy and (XAI) in healthcare systems is still in its early stages [70]. According to Arrieta *et al.* [71] explainable AI can be defined as: Given an audience, an XAI is one that produces details or reasons to make its functioning clear or easy to understand. Paniguttipani *et al.* [72] introduce Doctor XAI to deal with multi-labeled, sequential, ontology-linked data. Doctor XAI is a multi-label classifier that takes as input the clinical history of a patient in order to predict the next visit. In a most recent study, Loh *et al.* [73] reviewed XAI over the last decade and identified the following techniques that can be used for XAI:

1. SHapley Additive exPlanations (SHAP) that uses Shapley value from the concept of cooperative game theory.
2. Gradient-weighted Class Activation Mapping (GradCAM) technique which is an improved version of the original Class Activation Mapping (CAM) technique.

3. Local Interpretable Model-agnostic Explanations (LIME) that provides explanations in the form of the top significant features relevant for prediction

2.3 ML and DL for Diabetes

In this section we review some latest work involving the use of ML and DL in the field of Diabetes. Kumari *et al.* [74] focuses on enhancing the accuracy of diabetes detection by proposing an ensemble of machine learning algorithms using a soft voting classifier for binary classification into diabetes positive or negative. The process begins with data pre-processing, including normalization and label encoding, followed by data augmentation. The study uses the Pima Indian Dataset, consisting of specific health metrics for 768 individuals, with a mix of diabetic and non-diabetic patients. The models receive shuffled data points, and after individual predictions, the final prediction emerges from majority voting, leading to a more accurate and reliable outcome. Iparraguirre–Villanueva *et al.* [75] also worked towards the objective of early detection. Their research evaluates five different machine learning models: K-nearest neighbor (K-NN), Bernoulli Naïve Bayes (BNB), decision tree (DT), logistic regression (LR), and support vector machine (SVM), applying them to the same Pima Indian Dataset. In the study by Laila *et al.* [76] diabetes data consisting of 17 attributes were sourced from the UCI repository, specifically from clinical treatment records of patients from Sylhet Diabetes Hospital in Bangladesh. These data, validated by professionals, were utilized to assess the precision of predictions made using ensemble techniques. The research employed 520 instances, with each instance comprising 17 attributes used to predict the likelihood of diabetes. Each attribute, such as Age, Gender, Polyuria, and others, is detailed with specific values, providing a comprehensive dataset for analysis. The models used in this study include AdaBoost, Bagging, and Random Forest.
Jena *et al.* [77] uses DR images and DL to propose a novel approach of detecting DR. The screening technique introduced involves a unique asymmetric deep learning feature, utilizing U-Net for detailed segmentation of optic discs and blood vessels in the eye. Following this, the system employs a combination of a convolutional neural network (CNN) and a support vector machine (SVM) for classifying DR lesions. Lesions are categorized into four types: normal, microaneurysms, hemorrhages, and exudates. They use two public retinal image datasets, APTOS and MESSIDOR to demonstrate accuracy of their approach. Alkawid *et al.* [78] also studied DR. They used CNN, Inception-v3 and APTOS dataset as the previous research. However, they tried to detect five stages of DR instead of four. Thotad *et al.* [79] studied the application of deep learning algorithms to automatically differentiate between healthy skin and DFU-affected areas based on plantar thermograms. The effectiveness of the proposed model is benchmarked against existing deep learning frameworks, including DenseNet, VGGNet, and MatConvNet. They implemented the system using an Embedded GPU, demonstrating its compatibility with embedded systems, thanks to the NVIDIA Jetson

Nano toolkit. Yap *et al.* [80] summarizes experiments done on a dataset of 4,000 images (2,000 for training and 2,000 for testing) of DFU. Their work is a recap of DFU2020 competition where various sophisticated deep learning models, notably Faster R-CNN and its three distinct variations, YOLOv3, YOLOv5, EfficientNet were experimented. They concluded to the superior performance of a method based on Faster R-CNN which achieved the highest scores with a mean average precision (mAP) of 0.6940 and an F1-Score of 0.7434. Another conclusion from their study was that the application of ensemble methods could enhance the F1-Score, indicating a better balance of precision and recall.

2.4 Background and Preliminaries

2.4.1 Learning Nest

LN [8] focuses on knowledge construction by learners; "learning nests" correspond to group educational situations that take into account the specificities of individuals in terms of cultural specificities, academic level, preferences and values, personal and family life, environmental context, as well as the requirements of decision-making. The entire process takes place in a comfortable and non-conflictual space (the nest) that promotes the development of new knowledge and self-esteem, as well as the feeling of truly being able to act on health and illness (autonomy) within the social, cultural, and economic contexts of the patients. The Learning Nest has been deployed in diverse locations like Reunion Island, Mali, Mayotte, Mauritius, and Burundi [46]. Lifestyle interventions are promoted as the first tool, including physical activity and a healthy diet. The educators also focus on diabetes-associated complications like DFU, cardio-vascular disease, DR, chronic kidney diseases, and the need for periodic blood tests and check-ups. The idea to work towards an AI-powered ecosystem stemmed by our exposure to concept used for prevention, education and management of diabetes in relation to the LN. Figure 1.1 illustrates the LN concept.

2.4.2 Artificial Intelligence

AI is a sub-domain of computer science dedicated to the development of computer programs that strive to emulate human intelligence. This includes the capacity to analyze changing environments, make decisions, and learn from past experiences. It is often thought of being something new, but this has existed since long back and has now come to the forefront because of accessibility of huge data and processing power. Researchers from around the world have been increasingly drawn to Artificial Intelligence (AI) due to its cross-cutting applications across diverse domains such as agriculture, medical, economics, education, security, e-commerce, social networks, robotics and many more. Figure 2.2 shows the relationship between sub-domains of AI. Formal definition includes:

- AI refers to the way in which computer software emulates human cognitive processes. Like the human brain, an ANN (artificial neural network) is a network of interconnected layers that connects input and output signals [81].
- AI is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable [82].
- The text in the entries may be of any length.

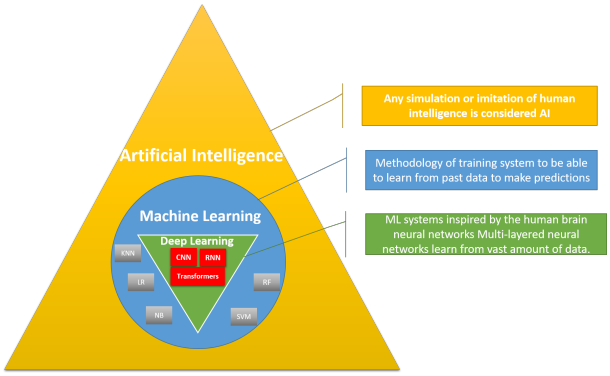


Figure 2.2: Relationship between Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL). ML includes algorithms such as Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Linear/Logistic Regression (LR), Naive-Bayes(NB) and DL includes architectures such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Transformers [83], [84], [85], [86].

2.4.3 Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) and computer science that focuses on the use of data and algorithms to model how humans learn, gradually increasing accuracy [87]. It employs minimal human intervention to analyze data and spot trends [88]. Machine Learning (ML) can be divided into three subtopics: supervised learning, unsupervised learning and reinforcement learning [89]. For the purpose of this thesis we concentrate essentially on supervised and unsupervised learning. Figure 2.3 give a graphical illustration of the process workflow for the training and deployment of a Machine Learning (ML) model.

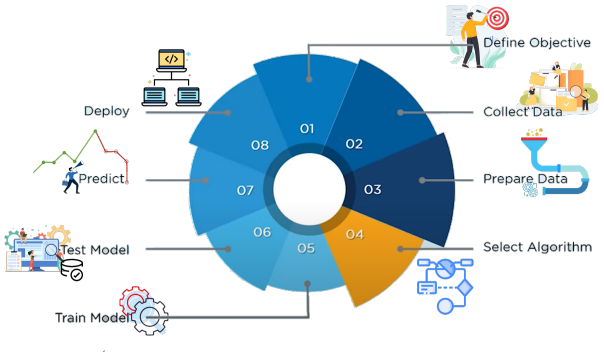


Figure 2.3: Graphical representation of the steps involved in Machine learning: 1. Define the objective. 2. Collect the required data. 3. Pre-processing data in the required format. 4. Select the ML algorithm appropriate for task in hand. 5. Train the model on Training data. 6. Test the trained model. 7. Use the model for prediction. 8. Deploy the model for use.

Supervised Learning

A machine learning task called supervised learning converts every input item to the required class label value. The data are referred to be labelled because it consists of pairs, inputs and its desired output. An object is mapped by the computer with the intended output after training [88]. The learner is required to learn (to approximate the behavior of) a function which maps a vector into one of several classes by looking at several input output examples of the function [90]. Over time, the learning algorithm refines its predictions of this output in an effort to narrow the gap between its predictions and the actual output [91].

Unsupervised Learning

As opposed to supervised learning, in unsupervised learning there are labeled data. The model is not fed inputs and expected labeled outputs. While supervised learning has gained more interest in recent year unsupervised learning is expected to become more important [92]. In unsupervised learning the system analyzes the unlabeled data to deduce a function that explains hidden patterns and writes observations from the dataset to discover these patterns [93].

Reinforcement Learning

Reinforcement learning (RL) is a type of machine learning in which an agent learns to perform a task through repeated interactions with a dynamic environment. One example where RL is used is in autonomous driving where we can see that the agent and environment are the basic components of reinforcement learning [94].

2.4.4 Deep Learning

Deep learning is a function of artificial intelligence (AI) that mimics the activity of the human brain to create patterns for information processing and decision making. Deep learning is a subset of machine learning in the field of artificial intelligence that can control learning from unstructured or networked data [87]. DL algorithms are characterized with powerful feature learning and expression capabilities compared with the traditional machine learning (ML) methods [95]. Deep learning is a powerful uncovers complex patterns within extensive datasets through the utilization of the back-propagation algorithm, enabling machines to adjust their internal parameters and compute representations across different layers based on the previous layer's representation [92]. The performance of deep learning models are dependent very large data which are necessarily always available in all field. This is actually one of the major issue that was faced during this thesis. Despite achieving human-like performance, AI models are still limited in their usage due to being perceived as black boxes, resulting in a lack of trust, which remains a primary reason for their limited practical application, particularly in critical field like healthcare [73]. Hence, there are many research concentrating on Explainable AI (XAI) which aims to provide a suite of machine learning techniques that enable human users to understand, appropriately trust, and produce more explainable models [96].

Training and learning Process of ANN

This section summarizes how an ANN learns and outputs the final model that is used.

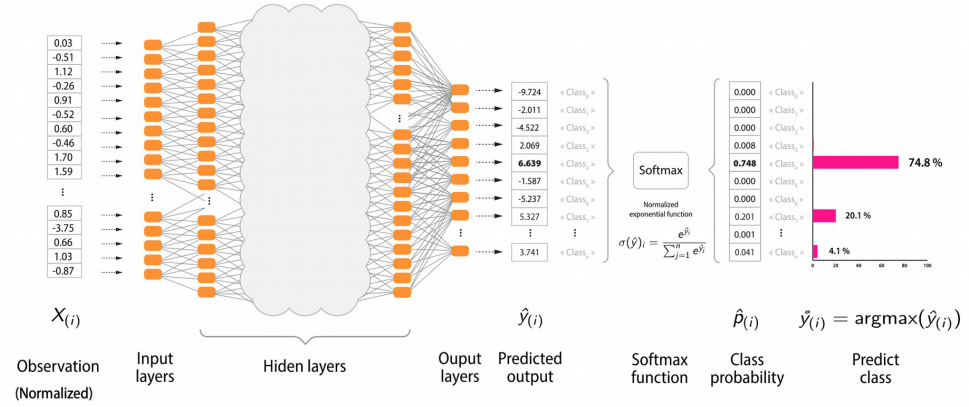


Figure 2.4: Graphical representation of an Artificial Neural Network showing the whole process for a classification problem [97].

From Figure 2.5, we can see that the learning takes place by feeding our observations to the network so as to obtain a prediction. We will then compare this prediction to the expected value. We will calculate the gradient of our loss function and update the values of our neural network. This process continues through a certain number of iterations, and each of these iterations is called an epoch.

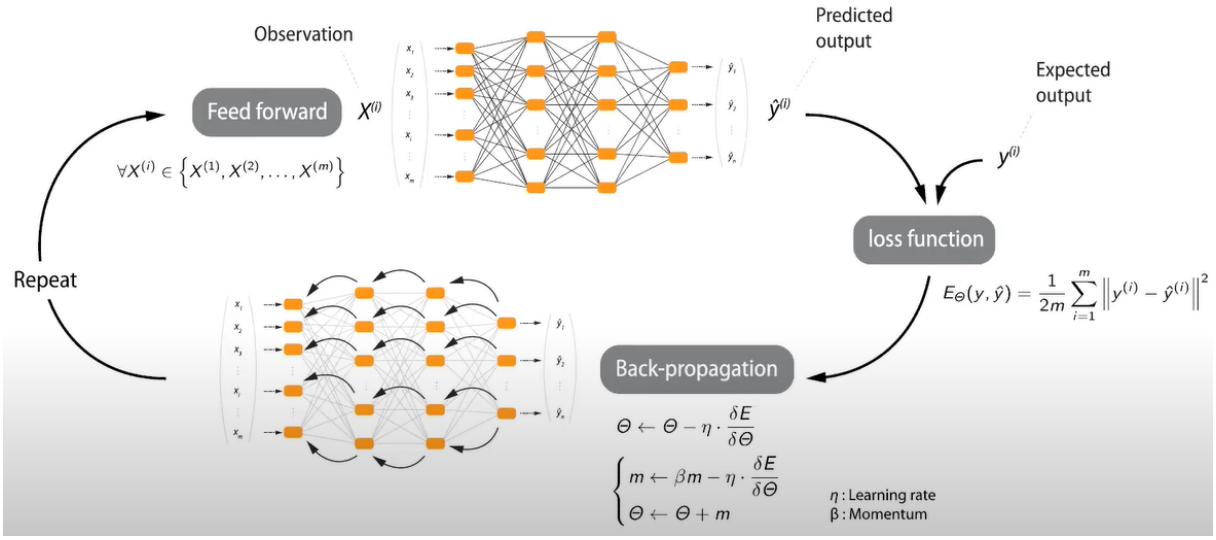


Figure 2.5: Graphical representation of an Artificial Neural Network showing the process of Feed-Forward, Back-propagation, Loss Function [97].

Algorithm 2.1 Training and Learning Process of an ANN

```
Load and preprocess training data and labels
Split data into training and validation sets
Initialize weights and biases randomly
Set learning rate and other hyperparameters
for epoch ← 1 to num_epochs do
    for i ← 1 to num_training_examples do
        input ← training data example i
        target ← true label for example i
        output ← forward_pass(input)
        loss ← calculate_loss(output, target)
        gradients ← backpropagate(loss)
        update_weights_and_biases(gradients, learning_rate)
    for j ← 1 to num_validation_examples do
        validation_input ← validation data example j
        validation_target ← true label for validation example j
        validation_output ← forward_pass(validation_input)
        validation_loss ← calculate_loss(validation_output, validation_target)
    average_validation_loss ← average of all validation losses
    Print("Epoch ", epoch, "/", num_epochs, ", Avg. Validation Loss: ",
    average_validation_loss)
```

2.4.5 Convolution Neural Networks

This section explains how the Convolutional Neural Networks (CNNs) work. Figure 2.6 shows all the layers of a basic CNN architecture.

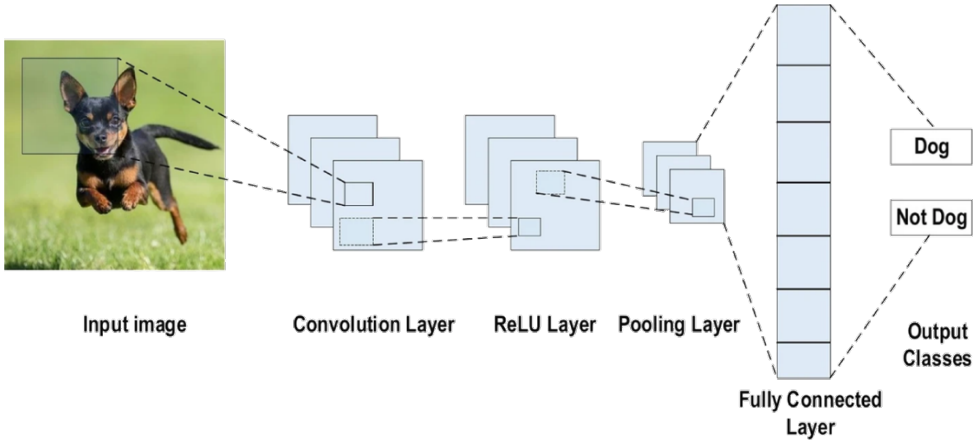


Figure 2.6: Schematic diagram of whole CNN architecture [98].

The name convolutional neural networks actually originated with the design of the LeNet [99]. It has since attained state of the art performance for various image related tasks using large image datasets [100], [101], [102]. The central element of this network revolves around the word **convolution**. It is thus important to understand this first before moving forward in the explanation. A convolution is actually a mathematical operation that takes two functions as input and produces a third function as output. When applied to the field of image processing, convolution is used to filter images. The filter is a small matrix of numbers that is used to modify the pixels of an image, and it is slid across the image to create another output image. The filter can also be called a convolutional matrix or the kernel. This is done to modify the initial image, for example, by blurring or edge detection.
The building blocks of the CNN is composed of several layers which individually explained.

Convolutional layer

The convolutional layer the central component in CNN architecture. It comprises a set of convolutional filters (also called kernel as explained above), also known as kernels. The input image, represented as N-dimensional tensors, undergoes convolution with these filters, resulting in the generation of the output feature map [98]. Figure 2.7 shows how the filter is applied to input image and calculated.

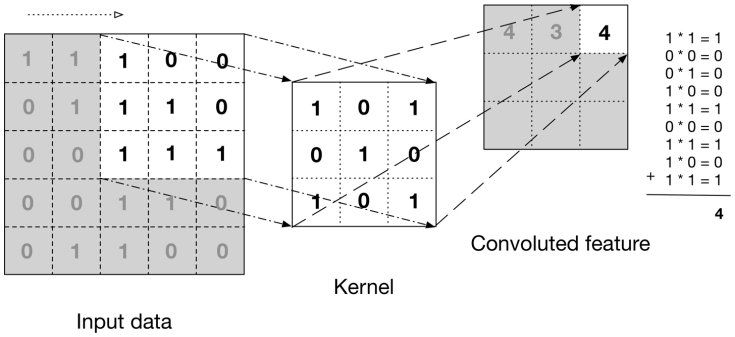


Figure 2.7: Illustration application of filter to input in a CNN [103].

It should be noted that two parameters are important, which are stride and padding which influences how the filter slides over the input image matrix. Padding is added to convolutional layers to prevent information on the borders from being lost. Stride determines the movement of the filter that iterates through an input image and is changed to control the density of convolution. Figure 2.8 illustrates the application of padding and stride.

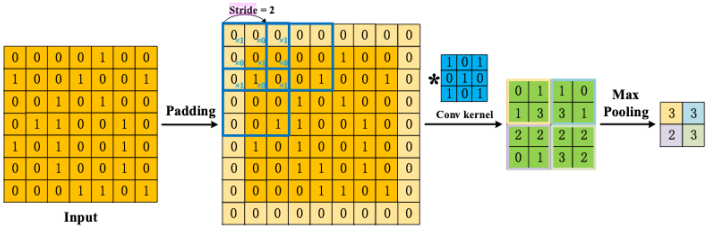


Figure 2.8: Illustration for Padding and Stride [104].

Pooling Layer

In a pooling layer takes as input the output from a convolutional layer and compresses it. The filter used in the pooling layer is smaller than the feature map and it takes a 2x2 square (patch) and reduces it to a single figure. For example, when using a 2x2 filter, the number of pixels in each feature map is reduced to one quarter of its original size. If the feature map is initially sized 10x10, the resulting output map would be 5x5.
Among the various functions can be employed for pooling the most common one are:

1. Maximum Pooling: This calculates the maximum value within each patch of the feature map.
2. Average Pooling: This computes the average value within each patch of the feature map.

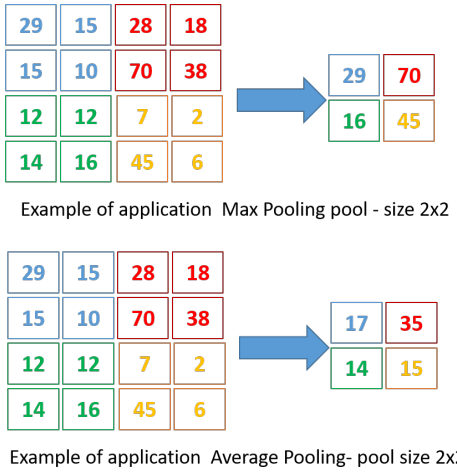


Figure 2.9: Illustration of Maximum pooling and Average pooling.

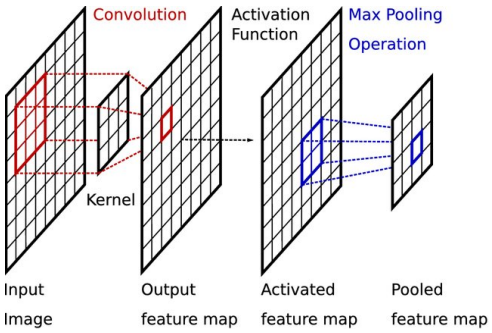


Figure 2.10: Illustration of the CNN architecture for image recognition [105].

Activation Layer

Activation layers are not technically "layers" (due to the fact that no parameters/weights are learned inside an activation layer) and are sometimes omitted from network architecture diagrams as it's assumed that an activation immediately follows a convolution. After each convolutional layer in a CNN, we apply a nonlinear activation function, such as ReLU or any of ReLU variants. We typically denote activation layers as RELU in network diagrams as since ReLU activations are most commonly used [106].

Fully Connected Layer

Neurons in Fully Connected Layers are fully connected to all activation in the previous layer, as is the standard for feedforward neural networks. This later is always places at the end of the network. Depending on the problem, an activation function can be added to promote the output for the network.

Hyperparameters

Hyperparameters are parameters whose values are set before starting the model training process. Optimizing hyperparameters in a deep neural network poses a significant challenge; nevertheless, identifying the optimal values for these hyperparameters leads to improved performance of the DL model [107].

2.4.6 Recurrent Neural Network

Recurrent Neural Network (RNN) are also known as sequence models that are used mainly in the field of natural language processing as well as some other areas such as speech-to-text translation and video activity monitoring. Unlike standard feedforward neural networks, recurrent networks retain a state that can represent information from an arbitrarily long context window [14]. Until recently, before the Transformer, they were widely used in natural language processing. While a fundamental assumption for Neural Networks (NNs) is the independence between successive inputs, this is not necessarily true in the real world. RNNs actually work by having a memory of what has been computed in the previous task. Figure 2.11 shows the simplified architecture of an RNN and its unfolded version.

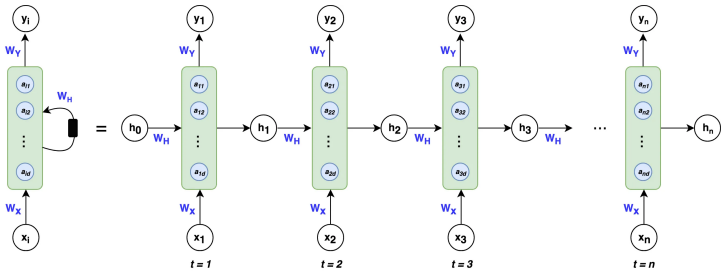


Figure 2.11: Illustration of the RNN architecture Natural Language processing [108].

Long Short-Term Memory

Long Short-Term Memory (LSTM) network is a type of an RNN [109] introduced to address problems of the aforementioned drawbacks of the RNN by adding additional interactions per module (or cell) making them earning long-term dependencies and remembering information for prolonged periods of time as a default [110].

2.4.7 Word Embeddings

As opposed to human which can read and understand words, it should be noted that that computers understands only 0 and 1. For machine learning algorithm it is not possible to process the words directly. It has to be converted into numbers which in turns are converted into vectors for mathematical operations. Word embedding analysis is a technique of natural

language processing that uses machine learning to train a neural network to predict the contexts within which words are used [111]. They are the representation of words such that they capture the semantic and syntactic relationships between words. They are represented as vectors in a high-dimensional space, where the proximity of two vectors reflects the semantic similarity of the words they represent.
There are different approaches for converting words to numbers such as the Word2Vec introduced by Mikolov *et al.* [112] which include also semantic meaning of the words. It uses unsupervised deep learning techniques to capture word associations from a large corpus of text leveraging both continuous bag of words as well as skip gram approaches to defining word context [113]. Figure 3.1 shows the model architecture based on CBODW and Skip-gram. Similar words when plotted in a vector space will be close to each other. In the initial paper, the authors propose the use of learned input embedding meaning that the token-to-vector conversion process is learned along with the main Machine Learning task.

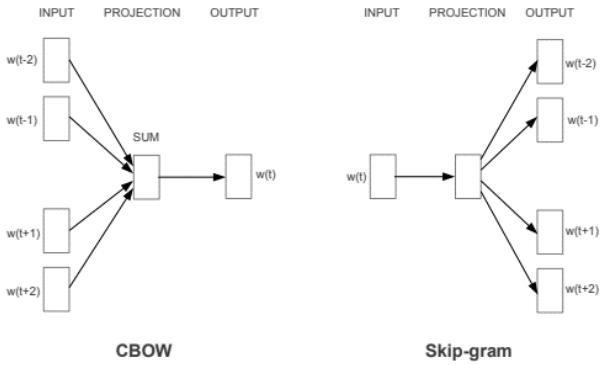


Figure 2.12: The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word [112].

2.4.8 Transformers

Prior to the prominence of Transformer architectures, Recurrent Neural Network RNN) and its variant Long Short-Term Memory LSTM were extensively employed for sequence-to-sequence tasks, particularly in Natural Language Processing (NLP). This thesis delves into the realm of NLP for processing diabetes-related news from digital mainstream media and for the development of intelligent conversational agents.
Transformers came to prominence from the now very famous article, "Attention Is All You Need" by Vaswani *et al.* [15]. It has taken the world of AI by storm with its powerful deep learning models. Notably, Transformer-based architectures have played a significant role in developing the renowned chatbot chatGPT, where GPT stands for Generative Pre-trained Transformer. This remarkable advancement in natural language processing showcases the transformative impact of Transformer models on AI applications. It overcomes the shortcomings of Recurrent Neural Network (RNN) as the architecture employs self-attention layers to process input

sequences simultaneously, enabling parallelization and faster training. They also introduced the concept of multiple heads of attention that were applied in parallel. The attention mechanism was introduced in 2014 [114]. It should be noted that transformers are now used in vision tasks as well, as introduced in the paper by Dosovitskiy *et al.* [115] which will be explained in section 2.4.9. Figure 2.13 gives the overall picture of the transformer architecture.

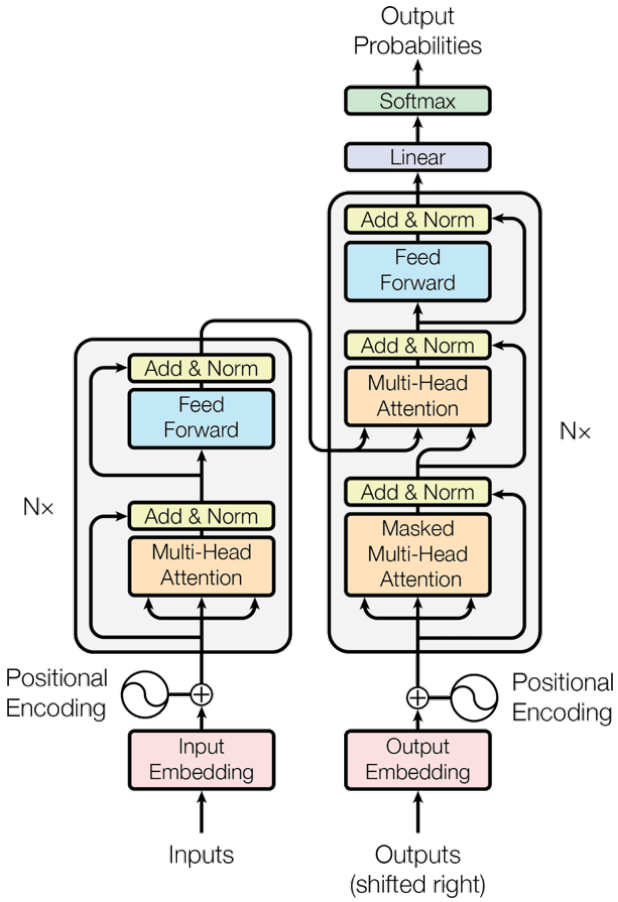


Figure 2.13: The overall architecture of the Transformer [15].

Input Embeddings

Transformer is a neural network. A neural network can only process words if they get converted to embedding representations. Figure 2.14 shows an example of the possible input embeddings for the sentence. The sentence must be converted into tokens, which are part of a fixed vocabulary and thereafter the tokens are converted in embedding vectors.
After that, tokens are converted into embedding vectors

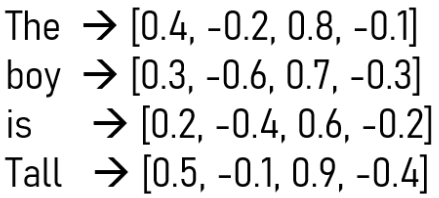


Figure 2.14: Example of input embeddings of the sentence "The boy is tall".

Positional Encoding

The position of a word in a sentence conveys meaning and using the same number of words but in different order completely changes the meaning of same as demonstrated by Figure 2.15.

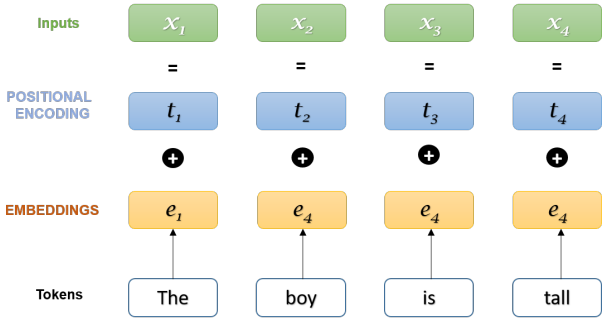


Figure 2.15: Example of input embeddings and Positional Encoding for of the sentence "The boy is tall".

We are processing the sequence all at once. We need know the position of tokens in the sequence. To address this problem, the transformer adds a positional encoding vector to each token embedding, obtaining a special embedding with positional information. Hence we need to The Position Encoding is computed independently of the input sequence. These are fixed values that depend only on the max length of the sequence. For instance,

- the first item is a constant code that indicates the first position
- the second item is a constant code that indicates the second position and so on.

These constants are computed using the formula below, where

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(\frac{pos}{10000^{(2i/d_{\text{model}})}}\right) \\ PE_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{(2i/d_{\text{model}})}}\right) \end{aligned} \tag{2.1}$$

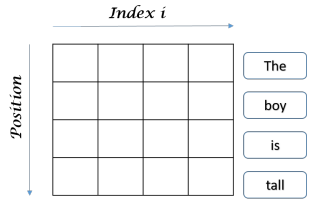


Figure 2.16: Illustration of the Vector with index and position.

- pos is the position of the word in the sequence
- d_model is the length of the encoding vector (same as the embedding vector)
- i is the index value into this vector.

Once the input embeddings are obtained position encoding is added to convey positional information in the input vector.

The inputs are ready to go through the encoder stack. There are 6 stacks in the original paper [15] as shown in Figure 2.17.

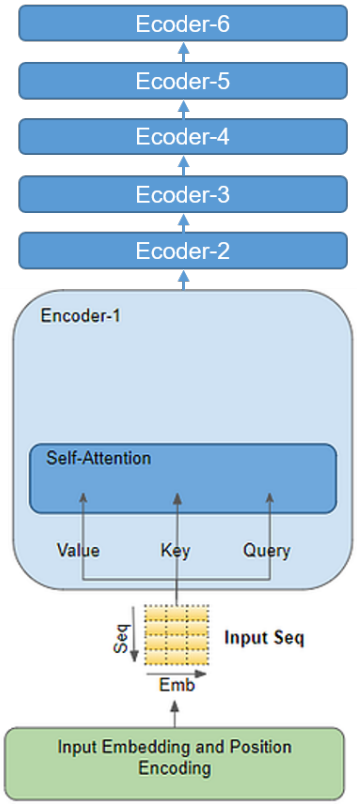


Figure 2.17: Illustration of the Encoder Stack inspired from [116].

Self-Attention

Self-attention is defined as the mechanism of relating different positions of a single sequence or sentence in order to gain a more vivid representation [117]. The Transformer architecture

uses self-attention by relating every word in the input sequence to every other word. We shall take an example to illustrate self-attention. While processing a word, Attention enables the model to focus on other words in the input that are closely related to that word.
We take the following example to illustrate Self-Attention:

- The cat drank the milk because **it** was hungry.
- The cat drank the milk because **it** was sweet.

In the first sentence, the word 'it' refers to 'cat', while in the second it refers to 'milk. When the model processes the word 'it', self-attention gives the model more information about its meaning so that it can associate 'it' with the correct word. This is shown in Figure 2.18

Figure 2.18: Illustration of Self-Attention Mechanism [118].

In the Attention layer of the encoder, the embedded sequence is passed through three Linear layers which produce three separate matrices - known as the Query, Key, and Value. These are the three matrices that are used to compute the Attention Score as shown in Figure 2.19.

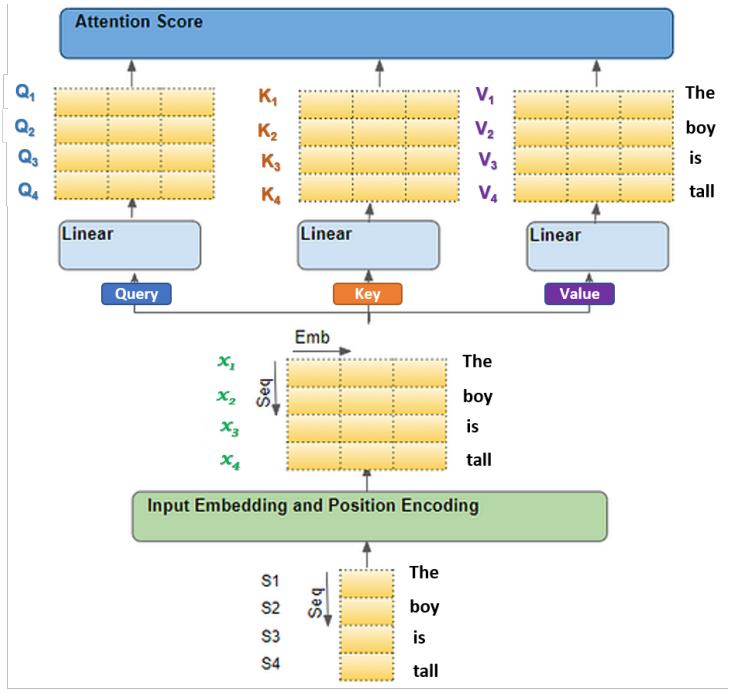


Figure 2.19: Illustration of calculation of Attention Score for Self-Attention. Inspired from [118].

The aim is to find the most important features input without the need to proceed the input sequentially one by one at a time. For this we generate three matrices using positional encoding aware matrix and multiplying it with another three independent linear layer as shown in the Figure 2.20. We create a query Q and a Key, K and use this to create an attention score using cosine similarity and illustrated in Figure 2.20

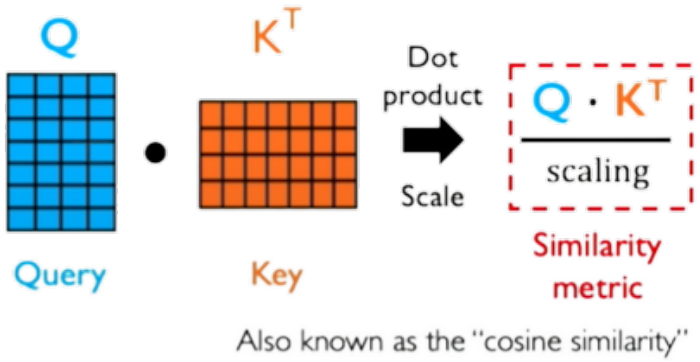


Figure 2.20: Illustration of calculating Cosine Similarity between Query and Key using the DOT operation [119].

The cosine similarity is eventually squashed between 0 and 1 using the softmax function. The results can be represented in the form of a heat map where it shows the relationship between each input word with each other word of the input sentence.

The final step of self-attention is to the take the generated attention weighting and multiply it with the Value matrix, V, and get output such that we get a feature matrix with reflects features that corresponds to high attention as shown in Figure 2.21. .

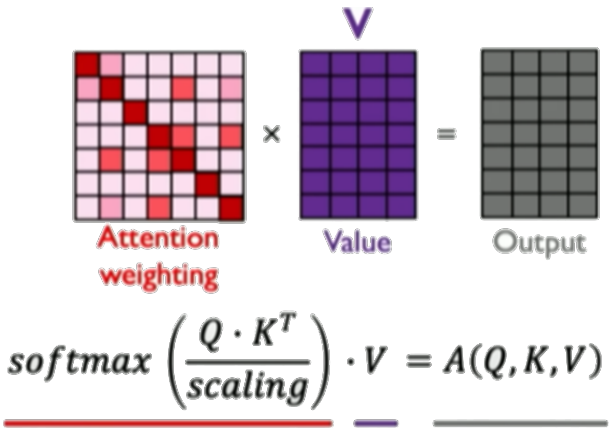


Figure 2.21: Final step for the calculation of the self-Attention [119].

The final equation is given as in equation 2.2 where Q, K, and V are the query, key, and value matrices, respectively. d_k represents the dimension of the key matrix, and softmax is the softmax function applied element-wise to the result of the inner product of Q and K^T. The final output of the self-attention mechanism is the element-wise product of the softmax result and the value matrix V.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

Decoder stack

The decoder is pretty much an encoder but with an additional encoder-decoder attention layer. Decoders pay attention only to the words before them, as opposed to encoders, which pay attention to every word regardless of order. As a result, the prediction for the word at the position, i, only depends on the words preceding it in the sequence. The block diagram of the decoder is shown in Figure 2.22.
The inputs of every decoder are:

- Previously generated sequence
- Encoder's output

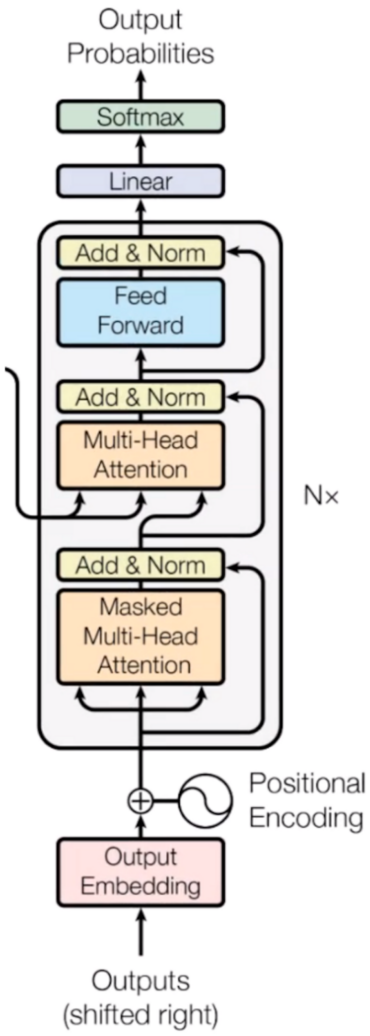


Figure 2.22: Block diagram of the decoder [15].

The detailed architecture of the transformer with comments is given in Figure 2.23

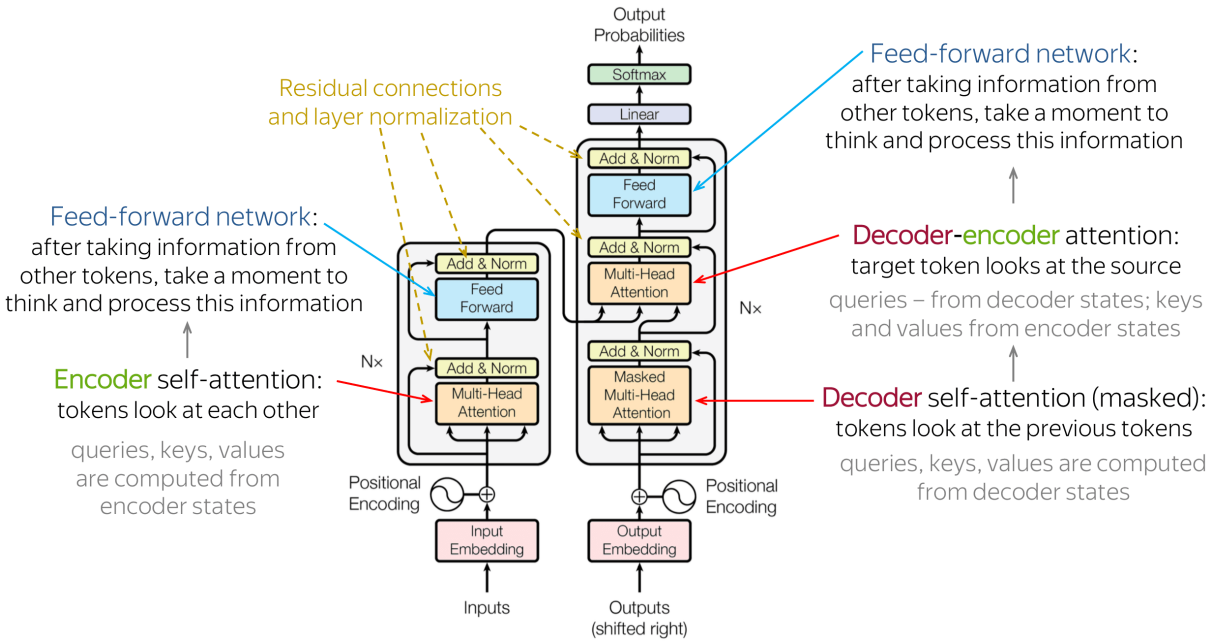


Figure 2.23: Summarized version of functions of various blocks that constitute the Transformer architecture [120].

2.4.9 Vision Transformers

Vision Transformers (ViTs) have emerged as cutting-edge and influential architectures in the field of computer vision, boasting remarkable potential to extract meaningful information from images. Their ability to unravel the complexities within images has positioned them as one of the most modern and dominant approaches in the domain of computer vision [121]. Vision Transformer (ViT) came to the forefront with the paper of Dosovitskiy *et al.* [115] "An image is Worth 16 x 16 Words". While Transformers where widely being used in the field of Natural Language Processing (NLP), it was not so prominent use with images where the use of Convolutional Neural Network (CNN) was exhibiting stat of the art performance. Vision Transformers (ViT) have recently achieved highly competitive performance in benchmarks for several computer vision applications, such as image classification, object detection, and semantic image segmentation. Figure 2.24 shows an overview of the Vision Image Transformer Model which is naturally inspired from the Transformer model of Vaswani *et al.* [15]. In Vision Transformers (ViTs), images are converted into sequences, enabling the models to predict class labels independently and learn image structures effectively. Each input image is treated as a sequence of patches, with each patch flattened into a single vector by concatenating the channels of its pixels. The resulting vectors are then linearly projected to achieve the desired input dimension. This approach allows ViTs to process images as sequences and capture their important features for classification tasks. The following summarises the steps involved when using Vision Transformers.

1. Split an image into patches
2. Flatten the patches
3. Produce lower-dimensional linear embeddings from the flattened patches
4. Add positional embeddings
5. Feed the sequence as an input to a standard transformer encoder
6. Pretrain the model with image labels (fully supervised on a huge dataset)
7. Finetune on the downstream dataset for image classification

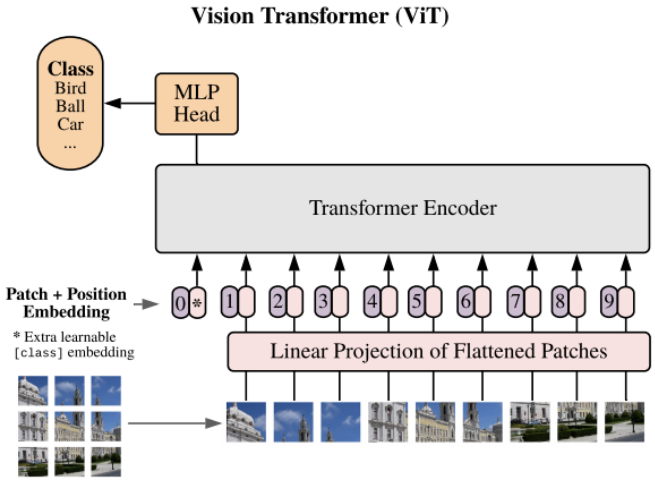


Figure 2.24: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence [115].

2.4.10 Generative Adversarial Networks

The concept of Generative Adversarial Networks (GANs) was initially presented by Goodfellow *et al.* [122]. In this framework, two models are trained simultaneously: a generative model G, which captures the underlying data distribution, and a discriminative model D, responsible for gauging the likelihood that a sample originates from the training data as opposed to G's generated output. The training objective for G involves maximizing the instances where D errs in its assessment. An analogy often invoked to elucidate this process draws parallels with a fraudster and a police officer. Generative models can be trained with missing data and can provide predictions on inputs that are missing data [123].

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.3}$$

2.4.11 Siamese Neural Network

A Siamese Neural Network is a category of neural network architectures that contain two or more identical subnetworks. By identical we mean they have same structure, same parameters and same weights. It was introduced by Bromley *et al.* [124] for signature verification written on tablet. During training the two sub-networks extract features from two signatures, while the joining neuron measures the distance between the two feature vectors. Verification consists of comparing an extracted feature vector with a stored feature vector of the person signing. These subnetworks which make up the siamese neural networks are constructed as feedforward perceptrons and utilize error backpropagation during the training process and they work in parallel comparing their output using the cosine distance as illustrated in Figure 2.25

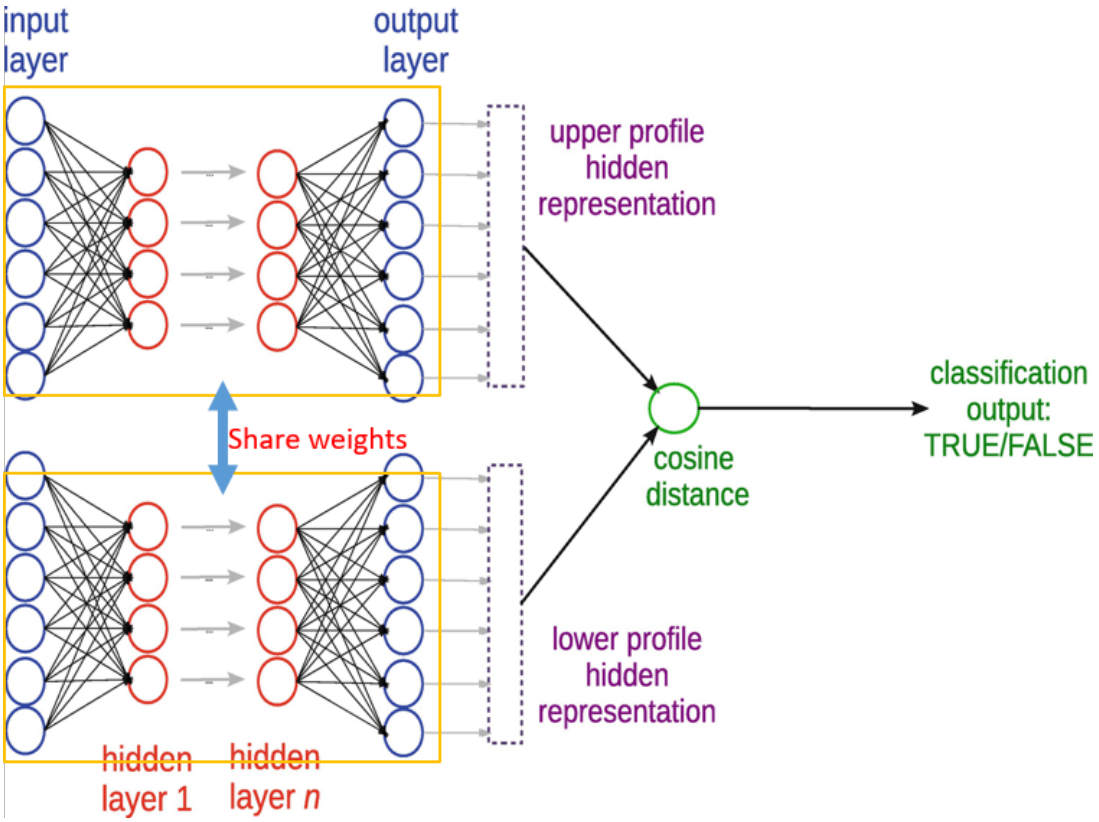


Figure 2.25: Representation of the structure of the siamese neural network model. The data are processed from left to right. The value of the cosine distance is a measure of the similarity between the input pair of data instances, as final output [125].

Chicco *et al.* [125] in their comprehensive study cites applications in the following field:

- Audio and signal processing
- Biology
- Chemistry and pharmacology
- Geometry and graphics

- Image analysis
- Medicine and health
- Optics and physics
- Text mining
- Video analysis

Deep Neural Networks (DNNs) are recognized for their reliance on extensive datasets for effective training. For instance, if a model is trained on 10 classes and an extra class is introduced later, the entire model necessitates retraining. In contrast, Siamese Neural Networks Deep Neural Networks (DNNs) are distinguished for their one-shot learning capability. This signifies that the incorporation of a new class does not mandate a complete retraining of the model. One-shot learning teaches the model to set its own assumptions about their similarities based on the minimal number of visuals. There can be only one image or a very limited number of them, in which case it is often called few-shot learning for each class.

As an example, consider differentiating between dogs and cats. A traditional ML model would necessitate a large dataset of thousands of training example [126], encompassing various angles, lighting conditions, and backgrounds. In contrast, one-shot learning defies the need for an extensive array of examples in each category. It harnesses its acquired knowledge from prior tasks of the same type, drawing connections among similar objects, and effectively categorizing unfamiliar objects into their respective classes.

During training of the SNN we need to ensure two input things:

1. The feature vectors of similar and dissimilar pairs should be descriptive, informative, and distinct enough from each other so that segregation can be learned effectively.
2. The feature vectors of similar image pairs should be similar enough, and those for dissimilar pairs should be dissimilar enough so that the model can quickly learn semantic similarity.

To ensure the model can learn similarity and dissimilarity it uses a loss function called Contrastive loss function. The contrastive loss function is a distance-based loss function that updates weights such that two similar feature vectors have a minimal Euclidean distance. In comparison, the distance is maximized between two different vectors. The contrastive loss function is give in 2.4

$$(1-Y)\frac{1}{2}(D_w)^2 + Y\frac{1}{2}(\max(0, m - D_w))^2 \tag{2.4}$$

In the equation 2.4, y represents whether or not the vectors are dissimilar, and D_w is the Euclidean distance between the vectors. When the vectors are dissimilar (y=1), the loss function minimizes the second term, for which D_w must be maximized (encourage more distance between dissimilar vectors). We want these vectors to have a distance of more than at least m (Which is a Margin), and we avoid computation if the vectors are already m units apart by defaulting to 0.

2.4.12 Transfer Learning

Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned [127]. It is based on using what has been learned from pre-trained models on a specific task and applying it to a different but related task. This avoids the need to start the learning process from zero. It thus reduces training time. Practically, it involves re-using state of a pre-trained model as the starting point as the starting point for the training of the model on a second task. Transfer learning is needed when there is a limited supply of target training data which could be due to the data being rare, the data being expensive to collect and label, or the data being inaccessible [128].

2.4.13 Language Model

While human learning how to communicated at an early age, computers cannot naturally grasp the abilities of understanding and communicating in the form of human language. This has been a challenge for researchers for a long time. With the use of AI it is now possible to achieve very good results. Language models are computational models that have the capability to understand and generate human language [44]. Language models (LMs) tries to model the likelihood of word sequences, so as to predict the probabilities of future or missing words. LLMs are systems that are trained on vast amounts of textual data and have the ability to generate human-like language and perform a wide range of language tasks such as translation, question-answering and sentiment analysis [129]. It is built on Transformer architecture. Large Language Models (LLMs) refer to Transformer language models that contain hundreds of billions (or more) of parameters and which are trained on massive text data [130].

BERT

When describing the Transformer model, translation tasks are frequently employed as examples to show the ground breaking effect. Since machine translation was known to be a very challenging task much research was focus on this and Vaswani *et al.* actually experimented their proposed transformer model on two machine translation tasks to demonstrate that the transformer model models performed well and in very less time [15]. It should however, be noted that time taken to train a Transformer model from scratch would take huge. Ideally there should be a trained Transformer based model that can be re-used for different task. Bidirectional Encoder Representation from Transformers (BERT) introduced by Devlin *et al.* [44]. BERT was built on the idea put forward by Generative Pre-trained Transformer (GPT) by pretraining a transformer model on a huge corpus of text and then fine-tuning it for specific NLP tasks [131]. Rashford *et al.* [131] demonstrated that large gains on tasks such that textual entailment, question answering, semantic similarity assessment, and document classification

can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task.
BERT use bidirectional transformer architecture stacking encoders from the original transformer on top of each other. This is illustrated in Figure 2.26

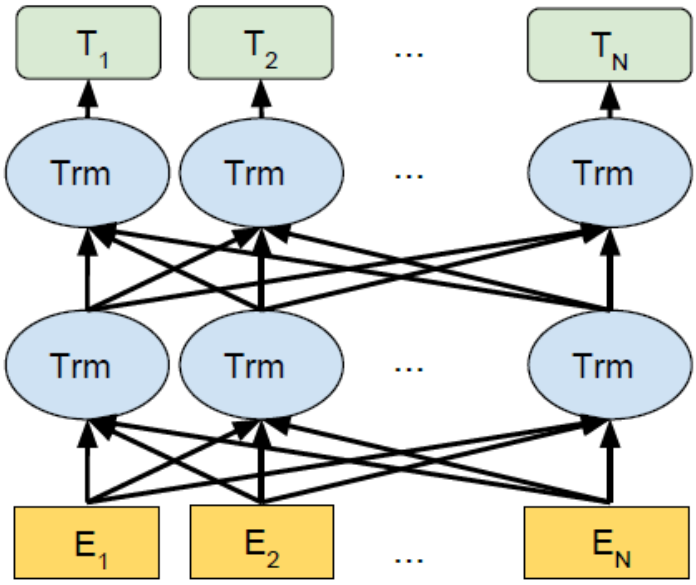


Figure 2.26: BERT uses many layers of bidirectional transformers adapted from [44].

BERT uses the encoder part of the Transformer, since it is goal is to create a model that performs a number of different NLP tasks. As a result, using the encoder enables BERT to encode the semantic and syntactic information in the embedding, which is needed for a wide range of tasks. BERT does not use the decoder part of the vanilla Transformer architecture. So, the output of BERT is an embedding, not a textual output. It takes the output of the encoder, and uses that with training layers which perform two innovative training techniques, masking and Next Sentence Prediction (NSP). Figure 2.27 shows how BERT computes input embeddings.

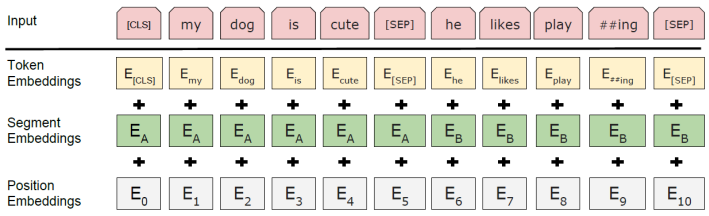


Figure 2.27: Input embeddings in BERT [44].

BERT was pre-trained on unlabeled data. The corpus for pre-training BERT had 3.3 billion words: 800M from BooksCorpus and 2500M from Wikipedia.

Generative Pre-trained Transformer

Generative Pre-trained Transformer (GPT) is a deep learning autoregressive language model (a simple feed-forward model) that produces human-like text from a set of words given in a specific context [129]. GPT is trained on a massive dataset of text and code. This allows it to learn the statistical relationships between words and phrases. As a result, GPT can generate text that is both coherent and grammatically correct. It is a type of large language model that uses deep learning to produce natural language texts based on a given input. It works such that when a user gives it an input, the generative pre-trained transformer creates a paragraph based on information extracted from publicly available datasets. GPT-3 developed by openAI is widely known for its use in CHATGPT. GPT was introduced in the year 2018 and used the BooksCorpus dataset to train the language model [131]. BooksCorpus had some 7000 unpublished books, which helped training the language model on unseen data. GPTs are machine learning algorithms that respond to input with human-like text. They have the following characteristics:

1. Generative. Given an Input they are able to generate new information.
2. Pre-trained. They first are trained in an unsupervised pre-training period using a large corpus of data. They are then fine-tuned in a supervised way to guide the model. Models can be fine-tuned to perform well in a specific task.
3. Transformers. The architecture is based on Transformer. They learn context by tracking relationships in sequential data. Specifically, GPTs track words or tokens in a sentence and predict the next word or token.

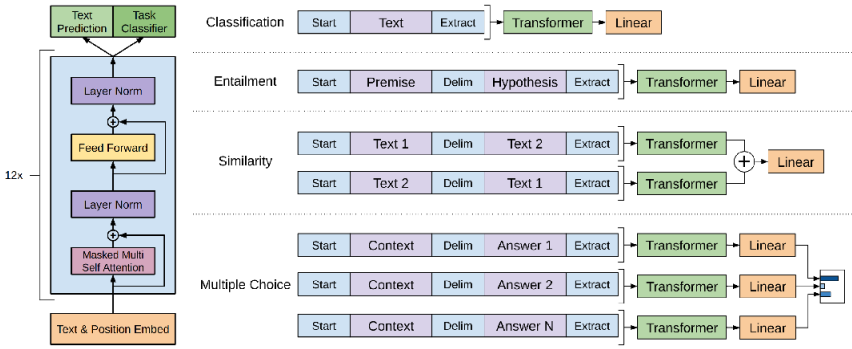


Figure 2.28: (Left)Transformer architecture and training objectives GPT. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer [131].

2.4.14 Ensemble Learning

Ensemble learning is a machine learning technique that combines the predictions of multiple models to improve the overall accuracy of the predictions. It flexibility and adaptability of ensemble methods and deep learning models have led to the proliferation of their application in bioinformatics research [132]. Ensemble learning algorithms typically work by training multiple

models on different subsets of the data or by training the same model multiple times with different hyper-parameter settings. The predictions of the individual models are then combined in some way to produce a final prediction. Mabrouk *et al.* [133] demonstrated the performance of an ensemble of several CNN pre-trained model and ViT for classification of pneumonia on chest X-ray images and their model outperformed state-of-the-art methods. Ensemble architectures perform classification task as a consequence of the voting performed by the individual models' accurate predictions [133].
Figure 2.29 shows how the voting is carried out in Ensemble Learning.

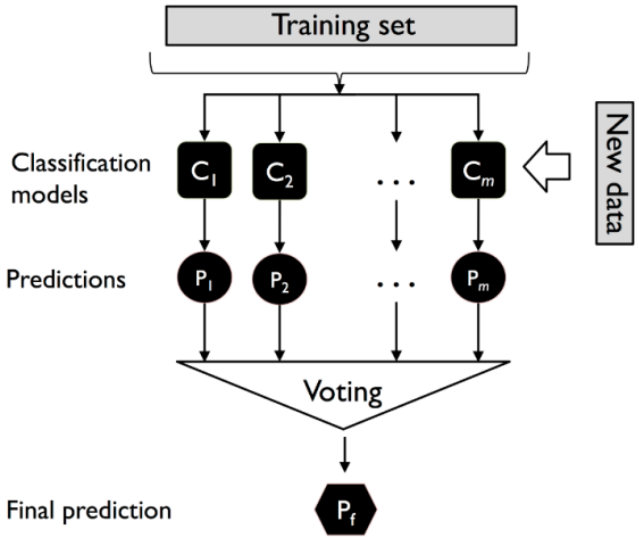


Figure 2.29: Illustration of voting in Ensemble Learning. C represents classification models and P represents prediction. Training data set is used to train different classifications models C1, C2,..., Cm. Then, new data are passed to each of the classification models to get the predictions. Finally, the majority voting is used for final prediction [134].

2.5 Evaluation Metrics

In Figure 2.3 where the Machine Learning (ML) process workflow is explained, two important steps are training and testing the model. The quality of the model can be subjectively evaluated as being good or bad. The model evaluation process is one of the most important steps in developing an effective ML model.
To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics: The evaluation metrics help us understand how well our model is performing with the available data. This approach enables us to enhance the model's performance by adjusting the hyper-parameters with the aim of achieving strong generalizations on unseen data.

In this thesis, image classification was a problem that was intensively investigated. It should be noted that performance metrics are not the same as loss functions. Classification metrics evaluate a model's performance and give you information on how good or bad the classification is.

2.5.1 Classification Problem Metrics

Confusion Matrix- Binary Classification

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. It shows the ground-truth labels versus model predictions as shown in Figure 2.30. Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class. Confusion Matrix is not exactly a performance metric but helps in the computation of other important metrics that researchers use to evaluate their models.

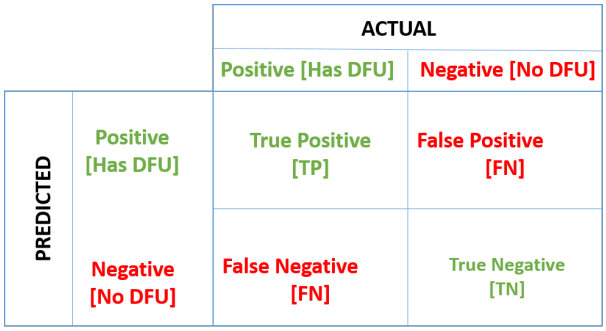


Figure 2.30: Example of Confusion Matrix for binary classification on presence of DFU or Not.

To understand the confusion matrix for this binary class classification problem of presence of DFU or Not, it is important to understand the following terms:

1. True Positive (TP) refers to a sample belonging to the positive class (Has DFU) being classified correctly.
2. True Negative (TN) refers to a sample belonging to the negative class (No DFU) being classified correctly.
3. False Positive (FP) refers to a sample belonging to the negative class (No DFU) but being classified wrongly as belonging to the positive class.
4. False Negative (FN) refers to a sample belonging to the positive class but (Has DFU) being classified wrongly as belonging to the negative class.

Confusion Matrix- Multi-class Classification

Columns represent the original or expected class distribution, and the rows represent the predicted or output distribution by the classifier as shown in Figure 2.31

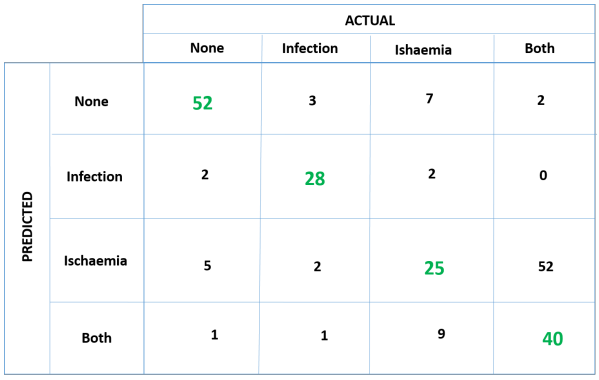


Figure 2.31: Example of Confusion Matrix for Multi-class classification for DFU.

The matrix is interpreted and can be converted into a one-vs-all type matrix (binary-class confusion matrix) for calculating class-wise metrics like accuracy, precision, recall.

Accuracy

The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions. It can be formulated as in equation (2.5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$

It should be noted that accuracy is not the best metric when we have imbalance data. Suppose the total number of images is 100 images, out of which 95 is healthy meaning no DFU and 5 has DFU. IF our model were to predict that all 100 is healthy, which is definitely not correct, we will still have an accuracy of 95%. In this case, our objective was to find a DFU case earliest possible and unfortunately despite an accuracy of 95% we will 100% of DFU. This goes to show we need to rely on other metrics also.

Precision

Precision metric is determined by dividing the number of correctly classified positive samples by the total number of samples classified as positive, including those that were classified incorrectly. This metric serves as an indicator of the model's ability to accurately classify samples as positive. The formula is shown in equation (2.6) where TP refer to the True positive and FP represent the False positive.

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

Precision does well in cases when you need to or can avoid False Negatives but cannot ignore False Positives.

Recall

Recall is derived by dividing the number of positive samples that were correctly classified as positive by the total number of positive samples in the dataset as illustrated in the equation (2.7). This metric is used to evaluate the model's capacity to identify positive samples accurately. Higher values of recall indicate that the model is better at detecting positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{2.7}$$

Recall, in contrast to Precision, is unaffected by the number of incorrect sample classifications. Additionally, Recall will be 1 if the model labels all positive data as positive.

F1-Score

F1-score is the harmonic mean of precision and recall values for a classification problem. F1-score symbolizes a high precision as well as high recall. It presents a good balance between precision and recall and gives good results on imbalanced classification problems.

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.8}$$

In classification problem F1-score favors classifiers with similar precision and recall. However, this may not always be the objective of researchers who are working on a model. In some cases, precision score is desired, while in others, recall score is preferred.
For example, if we train a model to detect videos that are safe for children, we would likely prefer a classifier that rejects many good videos (low recall) but only retains safe videos (high precision). On the contrary, let's say we train a classifier to identify shoplifters in surveillance images: we can settle for a classifier with a low precision of 30%, as long as its recall is 99% (in other words, security personnel will receive many false alarms (low precision), but almost all shoplifters will be intercepted (high recall). Hence depending on objective to be achieve we need to find a balance between Finding a balance between precision and recall.

Macro average F1-score

In multi-class classification with imbalanced data the main consideration will be Macro F1-Score. The formula is illustrated with the following equation(2.9) where n represents the number of class involved.

$$\text{Macro F1-score} = \frac{\sum_{i=1}^{n} F1\ score}{n} \tag{2.9}$$

2.5.2 Metrics in Natural Language Processing

No single metric suits every situation, just as no single machine learning model caters to all problems. Similar to how data collection and algorithm selection depend on the initial objective, the same principle applies to the choice of metrics. Specific metrics are better suited for particular types of applications.
Deciding on the efficiency of a model performance in a single metric is an inherently difficult task, and this is further complicated when applied to NLP domain because of structural and semantic of human language. In this section, we will introduce the crucial metrics that pertain to (Natural Language Processing (NLP)) tasks which were most relevant and considered in this research.

Exact Match

Exact match (EM) measures the proportion of documents where the predicted answer is identical to the correct answer. It follows a strict all-or-nothing metric; being off by a single character results in a score of 0. For example, for the annotated question answer pair "What is the capital of Mauritius?"- Answer: Port-Louis, a predicted answer in the form of "The capital is Port-Louis, would yield a zero score because it does not match the expected answer 100 percent.

F1-Score

While many NLP systems are evaluated using F1-score it should be noted that the latter lacks detail. For example, when two distinct problem achieve similar F-scores, they are not necessarily successful at the same kind of thing [135]. It is however, less strict than Exact Match. It measures the word overlap between the labeled and the predicted answer as illustrated in Figure 2.32.
To understand the how F1-score can be used in an NLP task we shall present tan example inspired by [136].

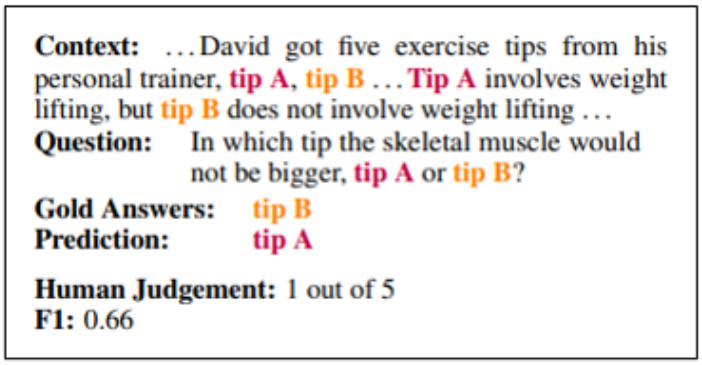


Figure 2.32: Example of F1-Score using Context, Question, Gold Answers [136].

BLEU Score

BLEU stands for Bilingual Evaluation Understudy. It is s a metric commonly used for evaluating the quality of machine-generated translations by comparing them to one or more human reference translations [137]. The BLEU score can range from 0 to 1, with 1 being the best possible score. A BLEU score of 0 means that the generated text does not match the reference translation at all, while a BLEU score of 1 means that the generated text is identical to the reference translation.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{n}\sum_{i=1}^{n} w_n \log p_n\right) \tag{2.10}$$

$$BP = \begin{cases} 1, & \text{if candidate length} > \text{reference length} \\ e^{(1-\frac{\text{reference length}}{\text{candidate length}})}, & \text{otherwise} \end{cases} \tag{2.11}$$

As evidence in [137] this metric was designed for machine translation evaluation tasks, however, it is also being used in other tasks such as text generation, paraphrase generation, and text summarization.

METEOR

METEOR stands for Metric for Evaluation of Translation with Explicit ORdering. It is also metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine produced translation and human-produced reference translations and was designed to explicitly address the weaknesses in BLEU identified above [138].
It creates an alignment by trying to map each token in a candidate to a token in a reference (and vice versa). A token is aligned to another token if they are the same, are synonyms, or their stems match. The alignment is aggregated into precision and recall values, which are combined into an F-measure score in which more weight is given to recall.

ROUGE

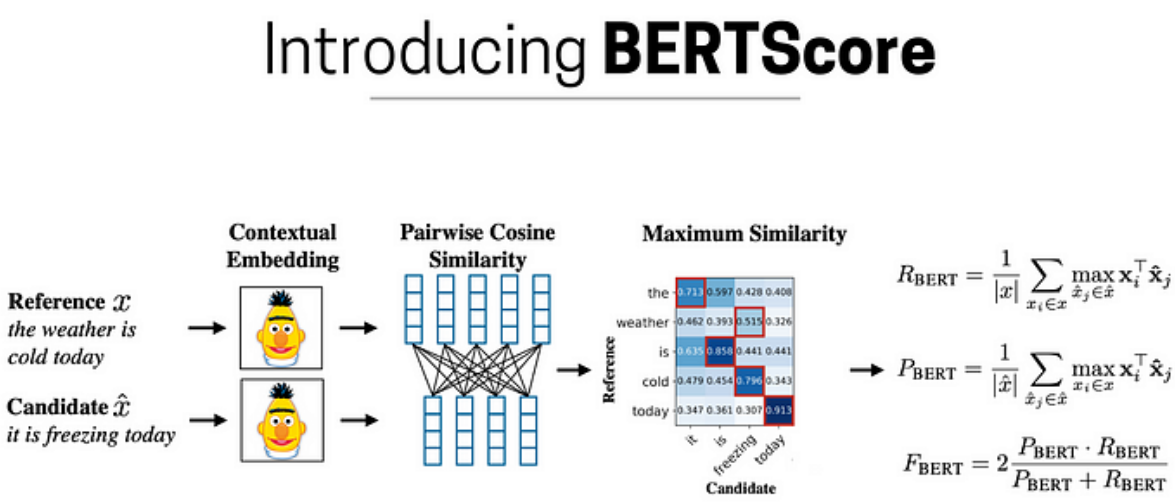
ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. It counts the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans [139]. ROUGE is actually a set of metrics.

1. ROUGE-N: It measures the number of matching 'n-grams' between model-generated text and a 'reference'.
2. ROUGE-L: It measures the longest common subsequence (LCS) between our model output and reference. All this means is that we count the longest sequence of tokens that is shared between both [140].

3. ROUGE-S: Stands for ROGE Skip-gram. It allows to ROUGE-S allows us to add a degree of leniency to n-gram matching and allows to search for consecutive words from the reference text, that appear in the model output but are separated by one-or-more other words [140].
4. ROUGE-W: Weighted Longest Common Subsequence. It evaluates the overlap between candidate and reference sentences, considering word weights. It prioritizes longer shared subsequences while considering the significance of individual words.

BERTScore

BERTScore is an automatic evaluation metric used for testing the performance of text generation systems. BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence and instead of exact matches, it computes token similarity using contextual embeddings [141]. As its name suggests it is based on BERT [44].



Source: Bertscore: Evaluating text generation with bert

Code for Bertscore is available at https://github.com/Tiiiger/bert_score

Figure 2.33: Illustration of the computation of the recall metric R_{BERT}. Given the reference $\hat{x}$ and candidate, BERT embeddings is computed and pairwise cosine similarity. Then highlight the greedy matching in red, and include the optional inverse document frequency importance weighting [141].

RQUGE

A key observation from the evaluation metrics discussed above pertains to the human-generated reference, often referred to as the "gold answer" in question answering. In this study, a crucial objective was to identify a metric capable of providing insight into whether the text generated

by our model would meet the desired standard or not. It is during this period that we came across the research by Mohammadshahi *et al.* [142]. RQUGE stands for Reference-Free Metric for Evaluating Question Generation by Answering the Question.

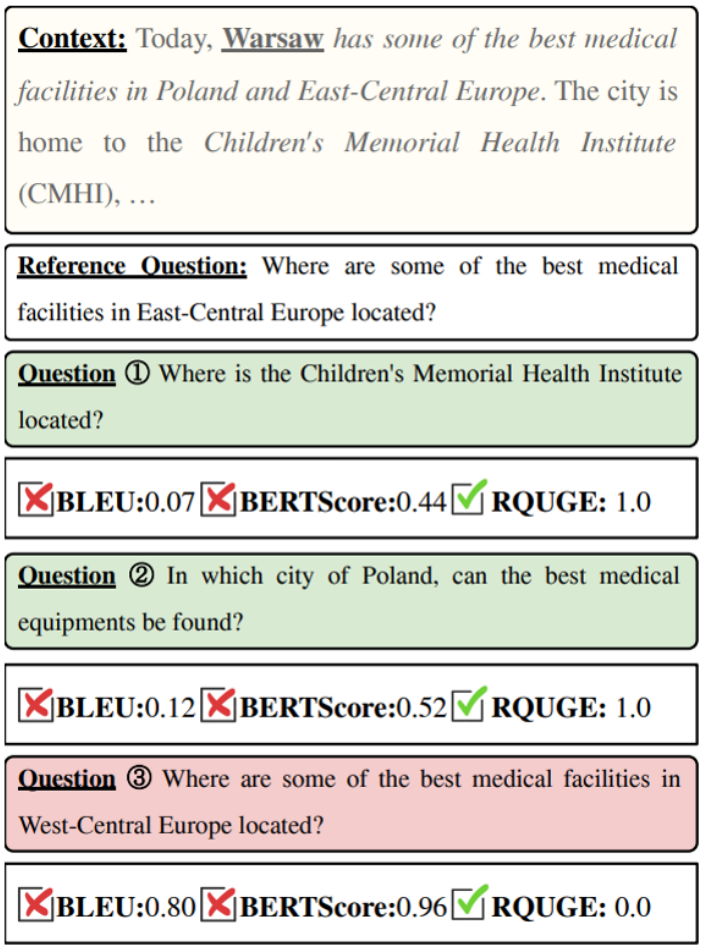


Figure 2.34: Normalised scores for different candidate questions. Metrics based on similarity to a reference question can penalise valid candidate questions, and compute a high score for unacceptable questions that are lexically similar to the reference. This can lead to the failure of reference-based metrics for valid questions, such as Q_1. Additionally, even paraphrases of the reference, like Q_2, may receive low scores. Furthermore, reference-based metrics may not detect small corruptions or variations in the reference, such as Q_3 [142].

RQUGE is a evaluation metric that can compute the quality of the candidate question without requiring a reference question. Given the corresponding context and answer span, our metric calculates the acceptability score by applying a general question-answering module, followed by a span scorer. Figure 2.34 demonstrates application of RQUGE compared to other metrics. Figure 2.35 shows a description of the RQUGE architecture applied to question answering.

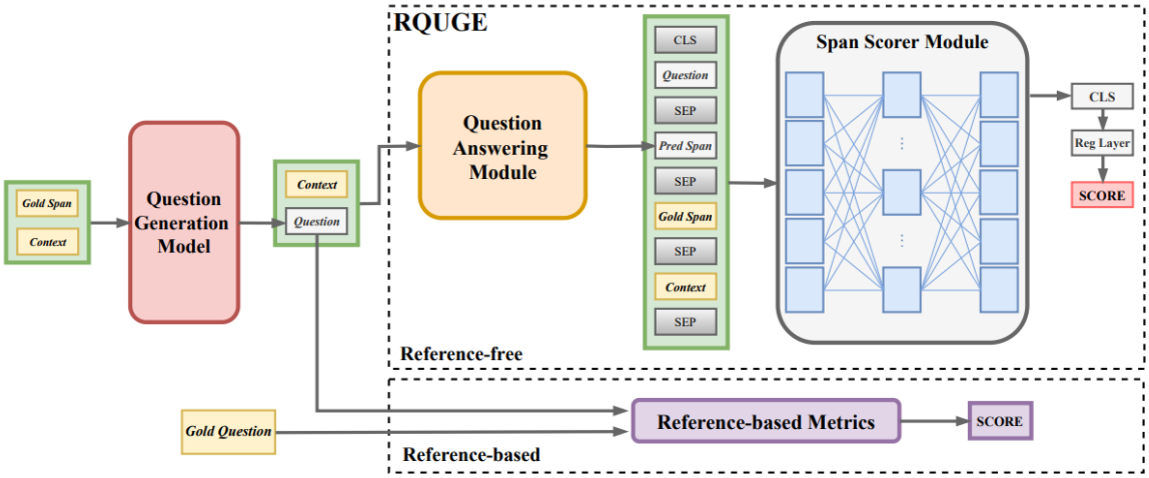


Figure 2.35: The architecture of RQUGE metric (upper-side) for the question generation task, which consists of a question answering and a span scorer module to compute the acceptability of the candidate question. Reference based metrics are also shown at bottom of the figure, where the score is calculated by comparing the gold and predicted questions [142].

3

Diabetic Foot Ulcer and Machine Learning

Summary

3.1 Introduction

Diabetic Foot Ulcer (DFU) is a devastating complication of diabetes that is associated with infection, amputation, and death and is affecting increasing numbers of patients with diabetes mellitus [143]. Mortality due to Diabetic Foot Ulcer (DFU) was high, with global mortality of diabetic foot ulcers standing at approximate 50% within 5 years [144]. When treating Diabetic Foot Ulcers, promptness and assertiveness can make a significant difference in slowing the wound's course and preventing the need for an amputation [145].

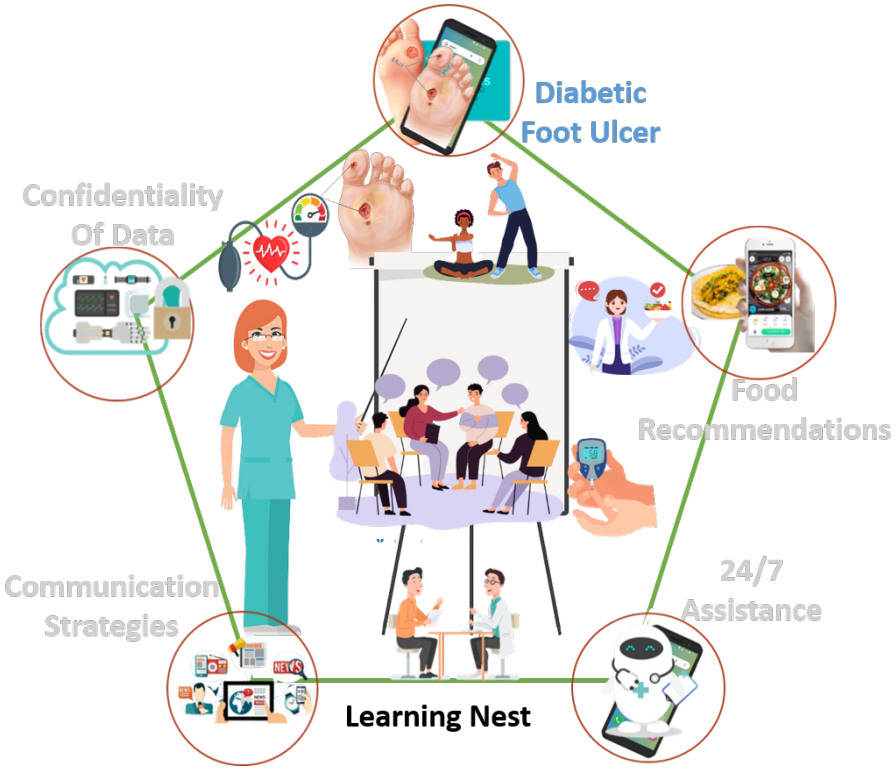


Figure 3.1: Objective is to investigate the use of AI for assisting in DFU care.

In this study, we explore the potential of leveraging state-of-the-art deep learning architectures for the early detection, prevention, and monitoring of treatment protocols pertaining to DFU. We investigate the following research question: RQ-1: How can deep learning be applied for

DFU management?. This forms part of our global objective of exploring the application of deep learning to the LN [46] as shown in Figure 3.1.

3.2 Diabetic Foot Ulcer - Classification

3.2.1 Motivation

In the current context of soaring demand for medical imaging and the prevailing challenge of staffing shortages in hospitals, the integration of AI tools holds promise as a potential solution [146]. With the increasing number of people either having diabetes or predicted to have diabetes, AI techniques can definitely contribute to reducing the progress, accompanying self-care, and boosting personalized care. According to the latest report published by International Diabetic Federation (IDF) [147], the prevalence of diabetic foot ulcers (DFU) varied mostly between 10.0% and 30.0%, and the prevalence of LLA, between 3.0% and 35.0%.
Currently, DFU is assessed by diabetes physicians and podiatrists in foot clinics and hospitals. There have been several research projects carried out based on the application of artificial intelligence and deep learning for DFU classification and even detection. This work is inspired by the successful implementation of AI in the medical field and takes advantage of the Diabetic Foot Ulcers Grand Challenge [148], which provides a labeled dataset with 4 classes that can be used by AI researchers to experiment and test the best model for classifications of DFU. In our endeavor to propose an AI-powered novel ecosystem, ultimately intended for integration with the Learning Nest, we chose to investigate Diabetic Foot Ulcer (DFU) and Artificial Intelligence (AI) techniques for the purpose of classifying Diabetic Foot Ulcer (DFU) into four distinct categories: non, infection, ischaemia, and both.

3.2.2 Related Work

In this section, we provide an in-depth review related to the main objective of this research, which is the classification of DFU images using AI techniques.
The study by Galdran *et al.* [149] compares the performance of CNNs and ViTs [115] for the classification of DFUs. The authors investigated the efficacy of the ResNeXt50 [150] architecture from Big Image Transfer (BiT) [151] and EfficientNet [152] for CNNs, as well as the ViT and Data-efficient Image Transformers (DeiT) [153]. In addition, they compared the optimization approaches of Stochastic Gradient Descent [154] and Sharpness-Aware Optimization (SAM) [155] for neural network training. The authors employed various data augmentation techniques during training, including random rotations, horizontal/vertical flipping, and contrast/saturation/brightness adjustments. For testing, four versions of each image

were generated, and the predictions were averaged for improved accuracy. Based on the results, the authors found that all pre-trained models performed better with the SAM optimizer. Specifically, the ResNeXt50 architecture demonstrated the highest performance on the test data. Interestingly, the authors achieved the highest scores by combining predictions from both CNN architectures. Through a thorough analysis of the various models, it can be concluded that CNNs outperform ViTs for the task of DFU classification.

Bloch *et al.* [156] introduced a novel approach for DFU classification using an ensemble of EfficientNets combined with a semi-supervised training strategy incorporating pseudo-labeling [157]. They address the challenge of class imbalance in the dataset, by using Conditional Generative Adversarial Networks (GANs) [158] to generate synthetic DFU images. They utilized the pix2pixHD [159] framework for conditional image generation. The proposed pipeline consisted of three phases: baseline training, dataset extension, and extension training. In the baseline training phase, the best-performing models were combined into an ensemble model. The baseline model was then employed to train the GAN with pseudo-labeling for both unlabeled and test images. The resulting dataset was subsequently used to retrain the EfficientNet variants, and the best-performing models were merged for ensemble prediction. Notably, the proposed approach demonstrated improved performance of 55.80% compared to the work of Gladran *et al.* [149] 52.82% for ischaemia class F1-score.

In the study conducted by Ahsan *et al.* [160], the authors focused on investigating various CNN-based deep learning architectures for binary classification. Specifically, they evaluated the performance of AlexNet, VGG16/19, GoogLeNet, ResNet50.101, MobileNet, SqueezeNet, and DenseNet. Employing a fine-tuning approach, the authors conducted experiments and assessed the accuracy of each architecture. Notably, the results revealed that ResNet50 exhibited the highest accuracy among all the tested architectures. It is important to note that although this research is recent, it lacked comprehensive information regarding the hyperparameters used.

In their work, Goyal *et al.* [161] introduced an ensemble CNN model that leverages the power of Inception-V3, ResNet50, and InceptionResNetV2 architectures [162]–[164]. The model combines bottleneck features extracted from these CNNs. During the training phase, the authors employed a strategy where the weights of the initial layers in the pre-trained networks were frozen to capture common features such as edges and curves. Subsequently, the later layers were unfrozen to focus on learning dataset-specific features. The ensemble-CNN model used the combined bottleneck features as input for binary classification, employing the Support Vector Machine algorithm. Comparative analyses were conducted against traditional machine learning methods, including BayesNet, Random Forest, and CNN-only approaches. Notably, the proposed CNN ensemble model outperformed all traditional machine learning techniques and CNN-only models, showcasing its superior performance in binary classification tasks.

Santos *et al.* [165] presented DFU-VGG, an innovative approach for the classification of diabetic foot ulcers (DFUs). The authors employed the VGG-19 architecture as the backbone of their CNN. Notably, they introduced batch normalization after each convolutional block. The

performance of DFU-VGG was evaluated against fine-tuned versions of VGG-16, VGG-19, InceptionV3, ResNet50, DenseNet201, MobileNetV2, and EfficientNetB0 networks in their original configurations. In a separate study conducted by E. Santos *et al.* [166], an experiment was conducted to investigate the performance of an ensemble model comprising various combinations of VGG-16, VGG-19, InceptionV3, ResNet50, and DenseNet201 architectures. The outcome of the experiment showed that the ensemble model consisting of VGG-16, VGG-19, and DenseNet201 demonstrated the highest performance among the tested combinations. This research sheds light on the effectiveness of ensemble models in improving the classification performance of DFU classification. Thotad *et al.* [79], also proposed to use a fine-tuned CNN backbone based on EfficientNet [152].

Khandakar *et al.* [167] approaches DFU classification by combining a CNN-based backbone with traditional machine learning algorithms. The features of DFU are extracted using a pre-trained CNN model. Then unsupervised method of k-mean clustering is used for clustering the images into three categories, namely mild, moderate, and severe.

Qayyum *et al.* [168] experimented with CNN and ViT. They used transformer-based architectures that were originally trained on the ImageNet dataset. The different vision transformers are fine-tuned by adding a fully connected layer with feature size (3072×768), ReLU activation (ReLU), dropout layer for regularization, and another fully connected layer with feature size (768×4) at the end layer of the different pre-trained transformers. The features extracted from the last layer of multiple transformers are concatenated pair-wise and applied to a fully connected layer at the end to concatenate the features of individual transformers and then pass to the classifier layer. Table 3.1 provides a concise overview of the outcomes from the research related to machine learning and DFU classifications.

Table 3.1: Summary of related work for DFU and Machine Learning

Research Work	**Architecture**	**Type of Classifcation**
Thotad *et al.* [79]	EfficientNet	Binary Class Abnormal / Normal
F Santos *et al.* [165]	VGG-16, VGG-19, InceptionV3, ResNet50, DenseNet201, MobileNetV2 EfficientNetB0	Multi-Class None / Infection/ Isca-hemia / Both
E Santos *et al.* [166]	CNN Ensemble [VGG-16, VGG-19, InceptionV3, ResNet50, DenseNet201, MobileNetV2]	Multi-Class None / Infection/ Isca-hemia / Both
Galdran *et al.* [149]	CNN [BIT- ResNeXt50, EfficientNet] ViT[ViT-base ,DeiT-small]	Multi-Class None / Infection/ Is-chaemia / Both
Qayyum *et al.* [168]	ViT[vit_base_patch16_224]	Multi-Class None / Infection/ Is-chaemia / Both
Ahmed *et al.* [169]	EfficientNet B0-B6 Resnet-50	Multi-Class None / Infection/ Is-chaemia / Both
Bloch *et al.* [156]	EfficientNets B0, B1, B2 Pseudo-Labeling GAN	Multi-Class None / Infection/ Is-chaemia / Both
Ahsan *et al.* [160]	AlexNet, VGG16/19, GoogLeNet, ResNet50.101, MobileNet, SqueezeNet, and DenseNet	Binary Class Infection / Ischaemia
Khandakar *et al.* [167]	K-Mean Clustering CNN	Multi-Class Mild / Moderate / High
Goyal *et al.* [161]	Ensemble CNN Support Vector Machine	Binary Class Infection / Ischaemia

Based on the findings from above, it is evident that ensemble methods have demonstrated favorable outcomes, as has the use of CNN architectures for image-related tasks. Additionally, promising results have been observed with the application of transformer-based architectures for image classification. In light of this, our research will use these insights and introduce an innovative model for DFU classification, which will be elaborated further in the forthcoming Section 3.2.3.

3.2.3 Proposed Architecture

On the basis of findings from our study of related works we propose an innovative architecture for the classification of images of DFU diseases, as illustrated in Figure 3.2. We use an ensemble of CNN and ViT as the backbone for the two "identical twins" networks of the SNN for feature extraction. A detailed explanation of the model training process can be found in Section 3.2.3. The training procedure involves utilizing DFU images and employing k-fold validation with K set to 5. A comprehensive description of the dataset used and pre-processing applied in this study is provided in Section 3.2.4. Once the model is trained, it becomes proficient in classifying input images into one of the four classes: none, infection, ischaemia, or both. To generate predictions on the test images, a K-Nearest Neighbors (KNN) model is integrated into the approach, conducting neighbourhood analysis to enhance the prediction of the model.

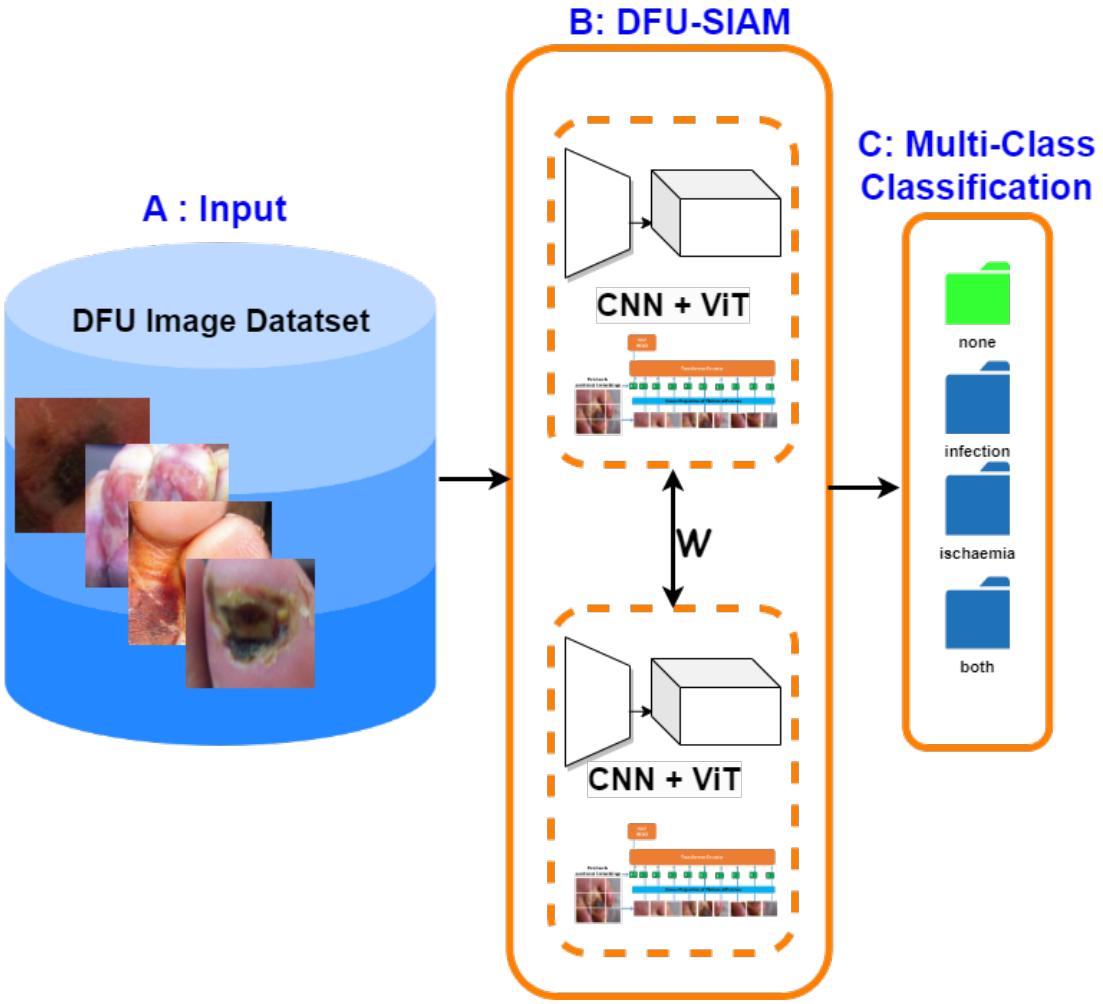


Figure 3.2: DFU-SIAM Architecture Overview for DFU Classification. A: Input images were sourced from the DFU2021 Dataset used for initial training and validation. B: The proposed Network, consisting of an ensemble of CNN and ViT within a Siamese Architecture. C: Visualization of the four distinct classes into which the DFU images are accurately classified.

The dataset used in this work was obtained from the DFUC2021 challenge [170], as detailed in Section 3.2.4. Due to the imbalanced nature of the dataset, various image augmentation techniques were employed to enhance the training process of the Siamese model. For both training and classification tasks, the KNN classifier was utilized.

DFU-SIAM

DFU-SIAM is a DFU disease classification model that implements a SNN. Figure 3.3 shows the ensemble model architecture. For the CNN backbone, we use EfficientNetV2S based on EfficientNet [152] architectures, which have been shown to significantly outperform other networks in classification tasks while having fewer parameters. EfficientNetV2S has fewer parameters,

making it more suitable for low-resource settings, and it uses a combination of efficient network design and compound scaling to achieve high accuracy with fewer parameters [171].

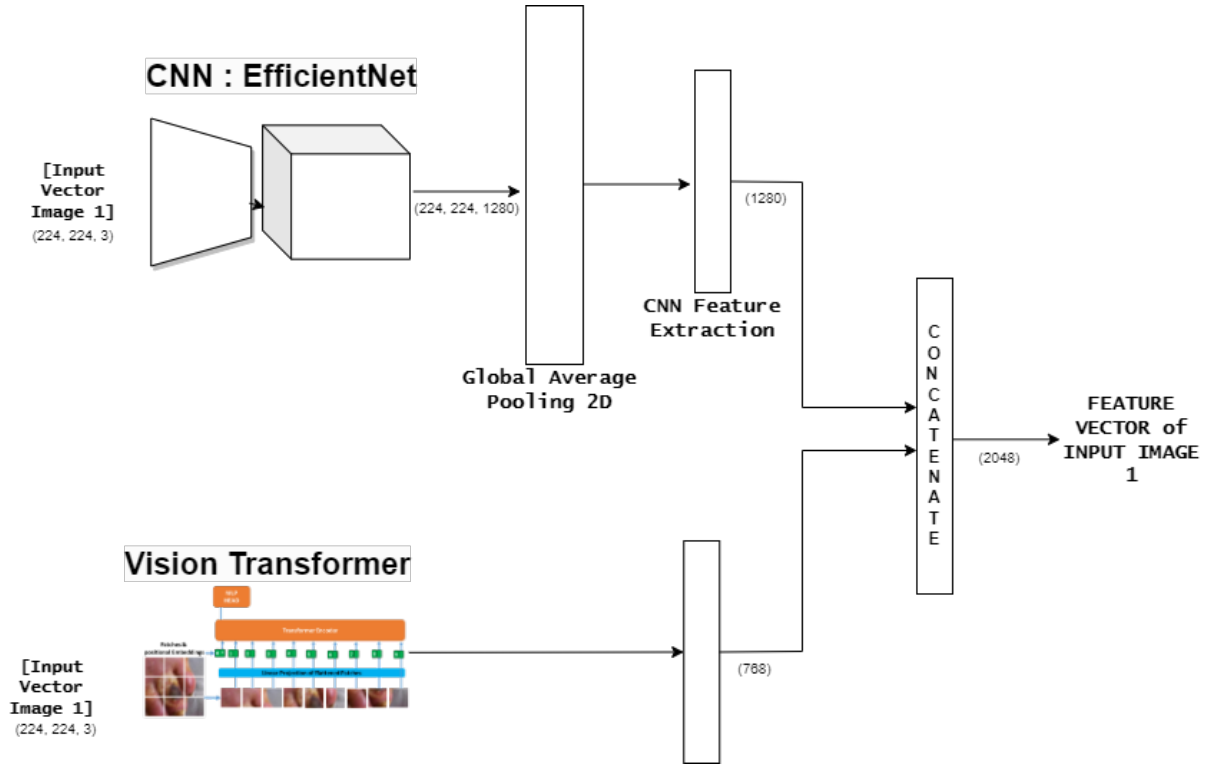


Figure 3.3: Block Diagram of the Ensemble Network, illustrating the internal architecture of the individual networks composing the SNN. The CNN utilized is EfficientNet, while the ViT employed is BEiT.

The second backbone of the ensemble model is based on ViTs, more specifically, Bidirectional Encoder representation from Image Transformers (BEiT). BEiT uses a pre-training task called masked image modeling (MIM) and stands for Bidirectional Encoder representation from Image Transformers, which draws inspiration from BERT [172]. MIM uses two views for each image, namely, image patches and visual tokens. The image is split into a grid of patches that are the input representation of the backbone Transformer. The image is "tokenized" into discrete visual tokens. During pre-training, some proportion of image patches are randomly masked, and the corrupted input is fed to Transformer. The model learns to recover the visual tokens of the original image instead of the raw pixels of masked patches.
The vector representation of image 1 is passed into both the EfficientNet model and the ViT model. In the EfficientNet, we remove the last dense layer from the pre-trained model to obtain the features from the last flattened layer (average pool). In the ViT model, we obtain the last hidden states, which contain all the patches from the last attention layer, except the classification token; then we flatten them and use another dense to reduce the shape to

make the output (features) have the same size as the feature extracted from EfficientNet. Finally, we merge the two feature sets.
Traditional Artificial Neural Networks learn by trying to minimise the loss function. Siamese Neural Network uses a different loss function, which is explained in the next section.

Loss Function of DFU-SIAM

While Siamese networks normally use contrastive loss, for DFU-SIAM we chose to implement Large Margin Cotangent Loss (LMCoT). Duong *et al.* [173] proposed LMCot as a novel approach for enhancing performance in verification and identification tasks. The LMCot loss utilizes the cotangent function instead of the cosine function. The cotangent function has a broader range of values, allowing for better optimization. Experimental results demonstrated that LMCot outperformed existing methods in various benchmark datasets and achieved state-of-the-art performance.
Once the chosen loss function has been established, it becomes crucial to outline the model evaluation and optimisation phase. In the subsequent section, we will provide a comprehensive explanation of how we intend to execute these steps to ensure optimal performance of the model.

DFU-SIAM Model evaluation and optimisation

This section explains the training, validation, and prediction processes of DFU-SIAM. The augmented dataset, consisting of training and validation images along with their respective labels, is loaded into the model. The model leverages the feature extraction capabilities of twin models to obtain the feature vectors of each image and employs the Large Margin Cotangent Loss as loss function. The objective of the learning process is to iteratively update the model parameters in order to minimize the distance between encoded features when the input images belong to similar classes while maximizing the distance when the input images belong to dissimilar classes. This ensures that the model learns to effectively discriminate between different classes by capturing meaningful patterns and representations in the encoded feature space. This overall process is detailed in Figure 3.4.

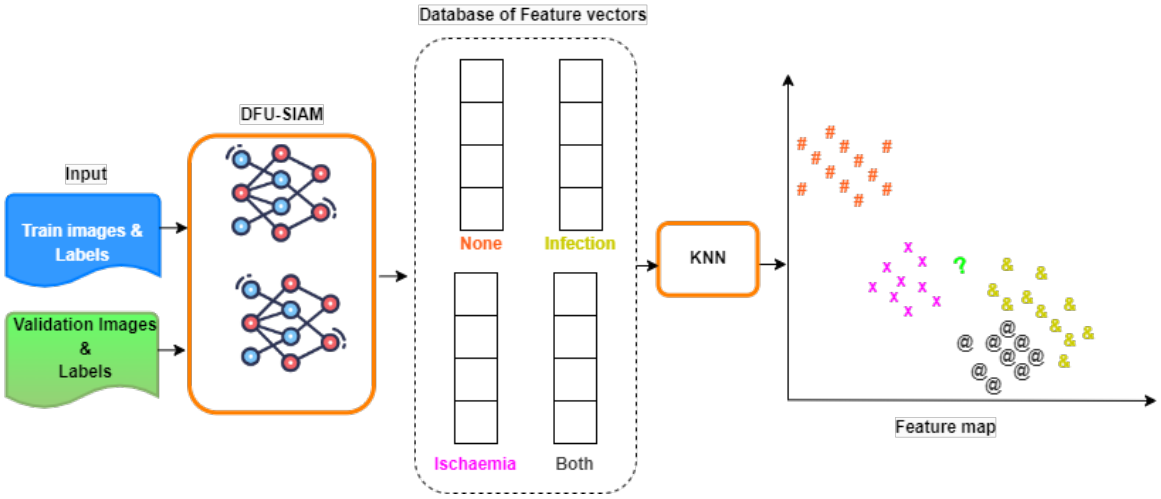


Figure 3.4: Illustration of the training process of DFU-SIAM, demonstrating the integrated approach of utilizing the SNN for feature extraction and machine learning for prediction during the training phase.

During the validation and prediction processes of the model, it is important to mention that the KNN classifier [174], is used as depicted in Figure 3.5. We iterate through values of K from 1 to 30 to determine the optimal value of K. The metric we use to select the best K is based on the Macro F1-score. KNN is a classifier model based on nearest neighborhood density estimation. For each epoch, an attempt is made to identify the optimal K value based on the Macro F1-score. Once the best K value is determined for an epoch, the corresponding weights are saved, and predictions are made on the test dataset. This iterative process ensures that the best predictions are obtained for each epoch of the test data.

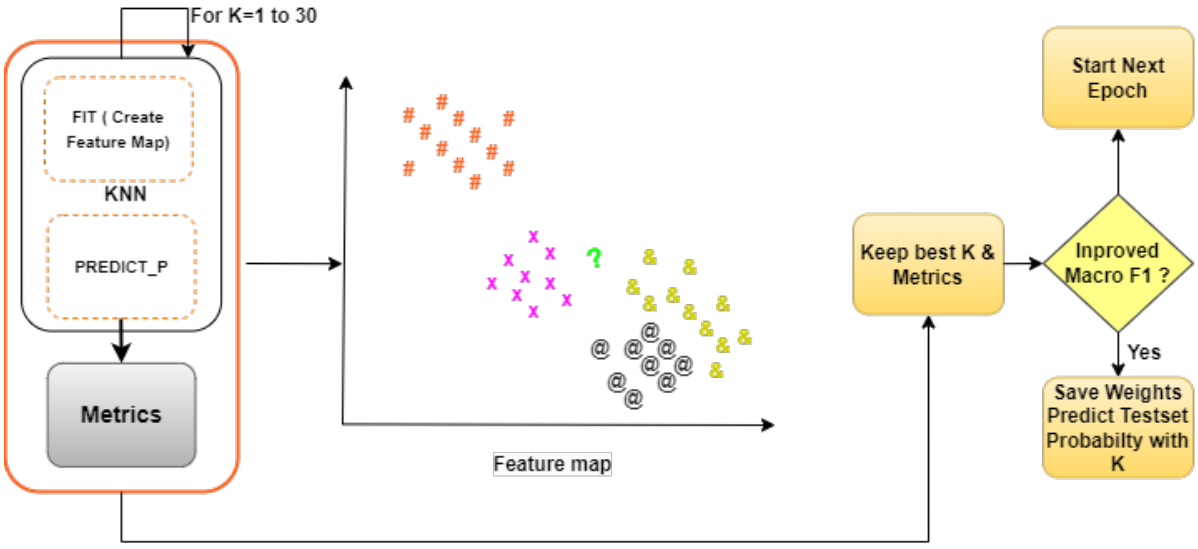


Figure 3.5: Schematic representation of the training process of DFU-SIAM including making predictions using machine learning algorithm KNN to determine the optimal value for K based on the highest Macro F1-score. The identified parameters are saved and subsequently employed for predictions on the test data.

The whole process of how DFU-SIAM classifies a test image is shown in Figure 3.6. The test image is fed into DFU-SIAM, which performs encoding and generates a compact feature vector within a lower-dimensional space. Within this reduced feature space, the encoded representation of the test image is compared to that of all the training samples using suitable distance measures. The classification is then carried out by employing the KNN algorithm.

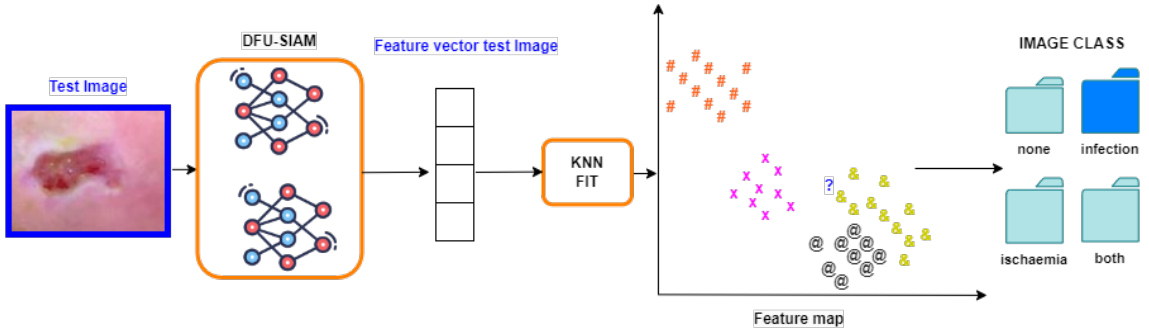


Figure 3.6: An overview of the application of DFU-SIAM for Test image classification. Input images are fed into DFU-SIAM, where they are encoded to generate feature vectors. The network then measures the distances between these feature vectors and all the training images. Utilizing the KNN algorithm, the predicted class for the input image is determined based on its proximity to the training samples.

3.2.4 Experimentation and Results

This section starts by providing a detailed description of the dataset utilized for the experimentation, including details on the preprocessing techniques employed. Additionally, the materials used in the experiments are outlined, and the results obtained are presented alongside a thorough comparison with relevant works in the field.
The quality of the dataset significantly influences the performance of deep learning models in terms of result accuracy. However, ethical reliability of the data source is equally important. In the next section, we will define the characteristics of the dataset employed in DFU-SIAM.

Dataset

In this section, we give an overview of the dataset we will use for this research.
Data quality is a crucial factor that directly affects the performance of supervised learning algorithms. The utilization of a representative and high-quality dataset is critical for achieving optimal accuracy and performance [175]. In this study, we obtained the dataset from the DFUC2021 challenge organized by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) society [170]. The proper licensing was also secured for this research, ensuring that all ethical and legal requirements were met.
Upon initial preprocessing, we observed that the dataset class distribution was imbalanced, with 621, 2555, 227, and 2552 instances belonging to the both, infection, ischaemia, and none categories, respectively, as shown in Figure 3.7. Such an imbalance poses a challenge to the performance of supervised learning algorithms, as they tend to be biased towards the majority class. To address this issue, we applied data augmentation techniques, as discussed in Section 3.2.4. It should be noted that Siamese networks, when combined with data augmentation techniques, can enhance the performance of various tasks. Data augmentation introduces variations to the training data.

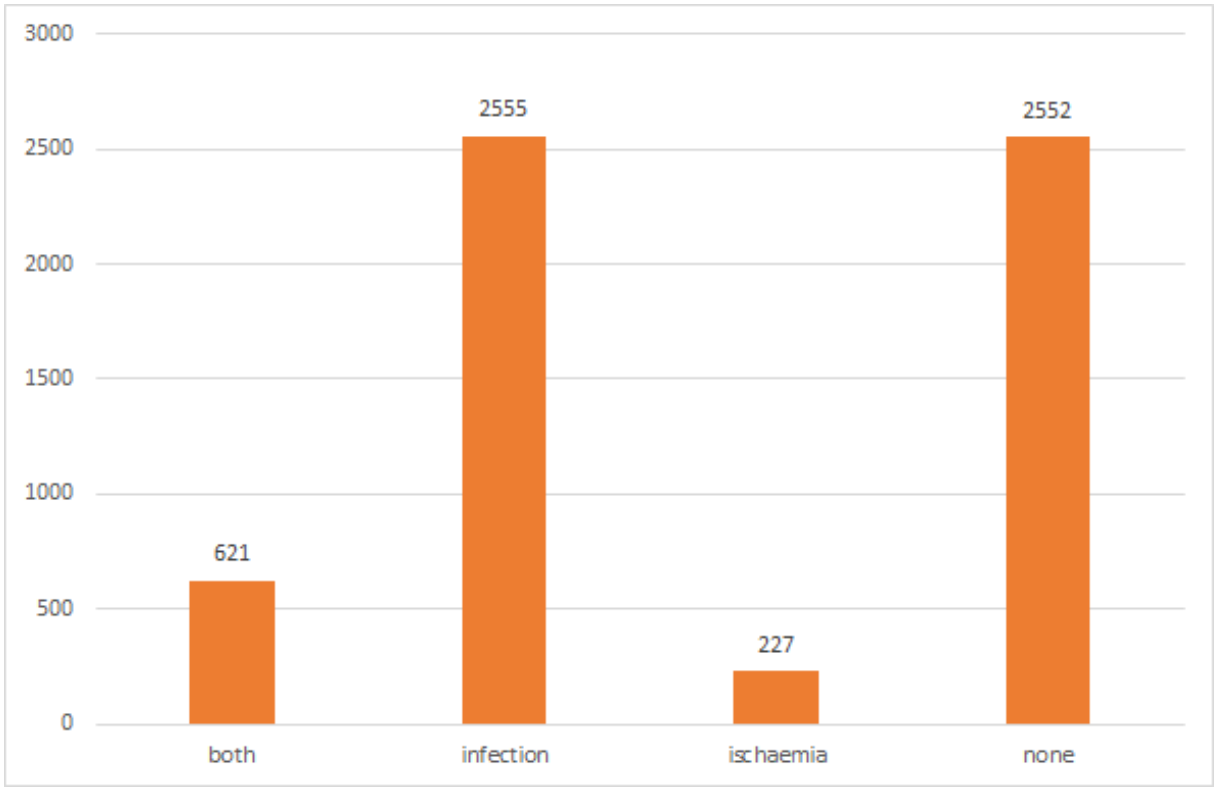
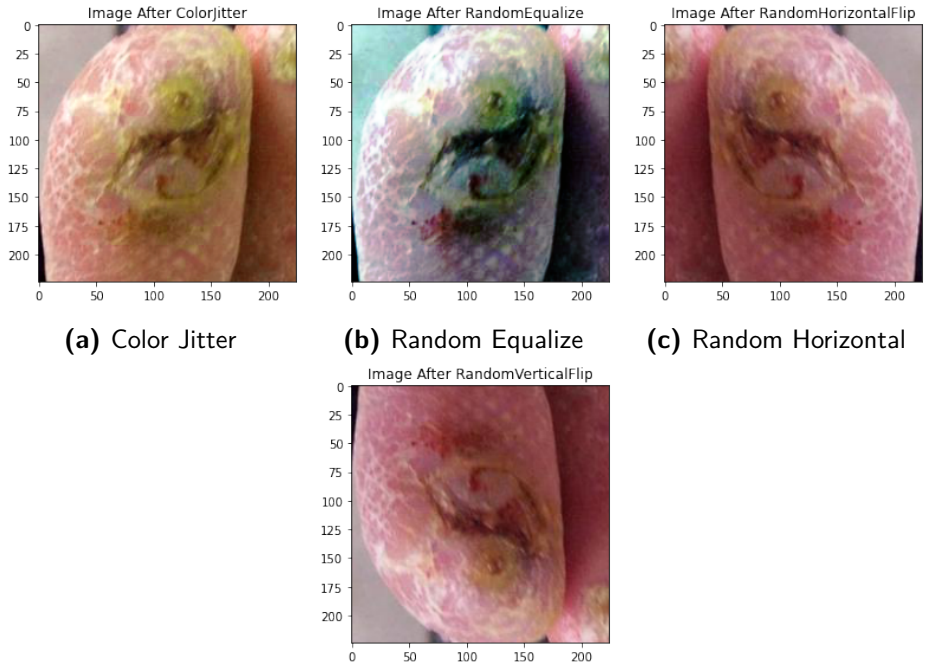


Figure 3.7: Class distribution of the DFU2021 Challenge dataset, illustrating the evident imbalance in the dataset.

Data Augmentation

Imbalanced data refers to a situation where one class of data examples has much more representation than the other classes [176]. The geometric transformations that were applied to our DFU dataset set images are illustrated in figure 3.8 and include:

- Colorjitter (brightness=0.1, contrast=0.1, Saturation=0.1, hue=0.1) Figure 3.8a
- RandomEqualize(p=0.2) Figure 3.8b
- RandomHorizontalFlip(p=0.2) Figure 3.8c
- RandomVerticalFlip(p=0.2) Figure 3.8d



(a) Color Jitter **(b)** Random Equalize **(c)** Random Horizontal

(d) Random Vertical

Figure 3.8: Demonstration of the application of geometric image transformations (a) Color Jitter, (b)Random Equalize, (c) Random Horizontal Flip, (d) Random Vertical Flip.

Prior to executing the model, it is most important to establish a suitable hardware and software setup, as they have an impact on the hyperparameters that will be employed. This setup is elaborated on below.

Experimental setup

The experimental setup was conducted on a Windows 10 Pro operating system running on a powerful hardware configuration comprising 64 GB of RAM and an Intel(R) Xeon(R) W-2155 CPU operating at 3.30 GHz. The system was further enhanced with an NVIDIA GeForce RTX 3060 GPU, boasting 12 GB of dedicated memory. To facilitate the experiments, the system was configured with CUDA version 11.7, Tensorflow 2.10.0, and Python 3.10.9.
The selection of hyperparameters in this study was influenced by the computational resources available. The batch size was set to 8, and the input images were resized to dimensions of 200 by 200 pixels with RGB channels. All models were run for 40 epochs. A fixed learning rate of 10^{-6} was employed. To optimize the parameters for prediction on the test data, the KNN algorithm was utilized. Additionally, test time augmentation (TTA) [177] techniques were applied to further enhance the prediction accuracy. TTA introduces random modifications to the test images, enabling the trained model to encounter augmented versions of the images multiple times. The predictions for each corresponding image were averaged, providing a more robust and reliable final prediction.

As explained in the previous section, our intention is to employ an ensemble of CNN and ViT as identical sub-networks of the SNN. The next section will explain the backbone that will be used.

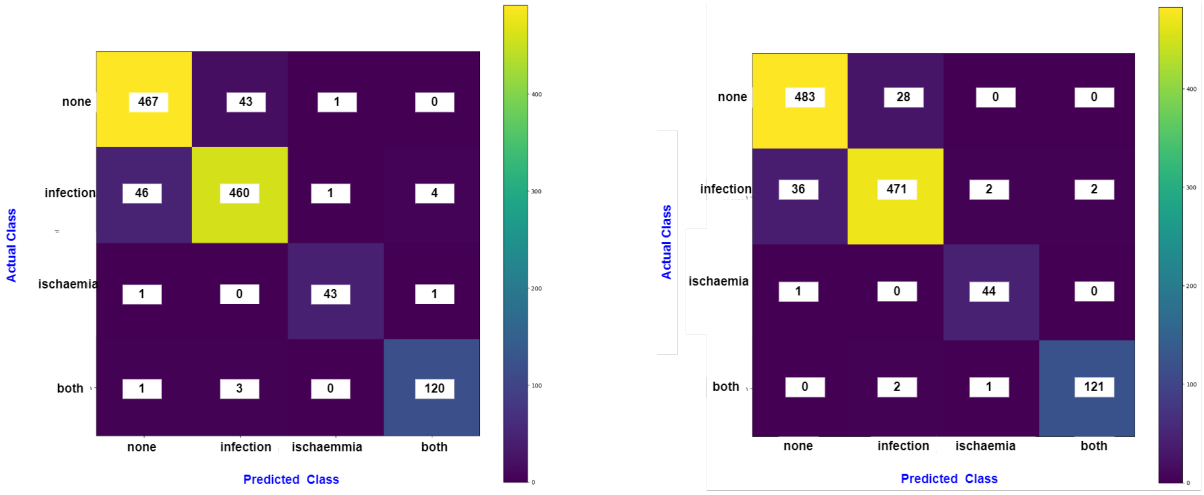
Experimental Strategy

For experimental strategy, we tested an ensemble of different combinations of CNN based and ViT based models. For the CNN, we maintained the EfficientNet. However, for the ViT we experimented with BEiT [172], [178] and SwinTiny(SwinT) [179]. Both will be tested and evaluated against related work.

Results

Confusion Metrics

The confusion matrix was obtained for the two variations of EfficientNet and ViT transformer, as shown in Figure 3.9. Figure 3.9a shows the confusion matrix when the backbone of our model is run with EfficientNet as the CNN backbone and SwinT(EfficientNet/SwinT) as the vision Transformer. Figure 3.9b shows the confusion matrix with EfficientNet and BeIT (EfficientNet/BEiT) as combined backbones. From the overall confusion matrix, the performance metrics are calculated. Table 3.2 and Table 3.3 show these metrics. By analysing the confusion matrices, we see that both models are wrongly predicting some instances of none class as infection and some as infection as none.



(a) CM: EfficientNet/SwinT **(b)** CM: EfficientNet/BEiT

Figure 3.9: Confusion matrix results obtained from applying two different ensembles as identical networks (a) Confusion Matrix with an ensemble of EfficientNet/SwinT. (b) Confusion Matrix with an ensemble of EfficientNet/BeiT.

Table 3.2: Metrics from Confusion matrix EfficientNet/SwinT

	Precision	**Recall**	**F1-score**
none	0.92	94	0.93
infection	0.93	0.91	0.92
ischaemia	0.96	1	0.98
both	0.98	0.98	0.98
accuracy			0.93
macro avg.	0.95	0.96	0.95
weighted avg.	0.93	0.93	0.93

Table 3.3: Metrics from Confusion matrix EfficientNet/BEiT

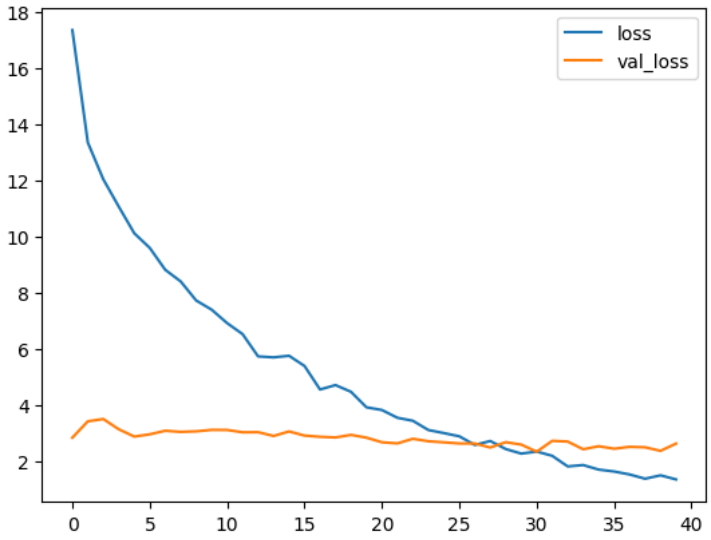
	Precision	**Recall**	**F1-score**
none	0.93	0.95	0.94
infection	0.94	0.92	0.93
ischaemia	0.94	0.98	0.96
both	0.98	0.98	0.98
accuracy			0.95
macro avg.	0.95	0.96	0.95
weighted avg.	0.94	0.94	0.94

From Table 3.2 and Table 3.3 we can see that EfficientNet/BEiT has a better accuracy of 95% compared to 93% of EfficientNet/SwinT. The Macro average F1-score is same at 0.95. EfficientNet and BEiT model has a higher Macro F1-score for the classes none and infection.

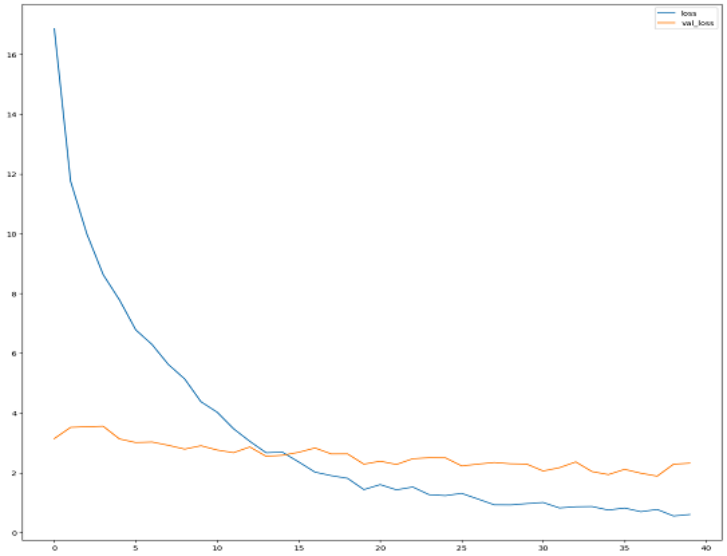
Loss

The loss function provides insights into the effectiveness of the models in minimizing errors and improving their predictive performance. By analyzing the loss curves, we can observe the behavior of the models over time and assess their training progress. As far as training loss is concerned, we can see in Figure 3.10 for both models that training loss decreases at a constant rate. This indicates effective learning and model improvement throughout the training process. Furthermore, by analysing validation loss curves, we can assess how well the models are learning and how effectively they are adapting to the validation dataset. The validation

loss very quickly stagnates for both models. However, we can witness a constant decrease for the EfficientNet and BEiT model as shown in Figure 3.10b.



(a) Loss curve of the model with an ensemble of EfficientNet and SwinT, trained for 40 epochs.



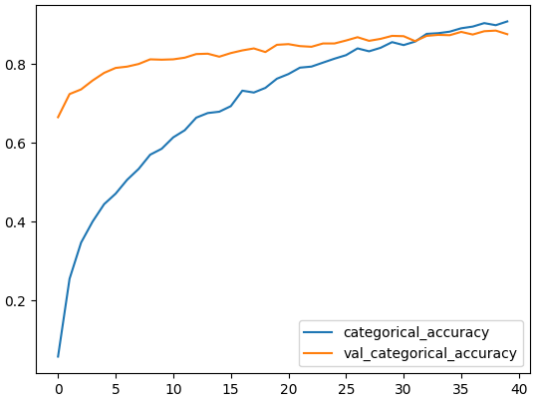
(b) Loss curve of the model with an ensemble of EfficientNet and BEiT, trained for 40 epochs.

Figure 3.10: Loss curves of the two models being experimented

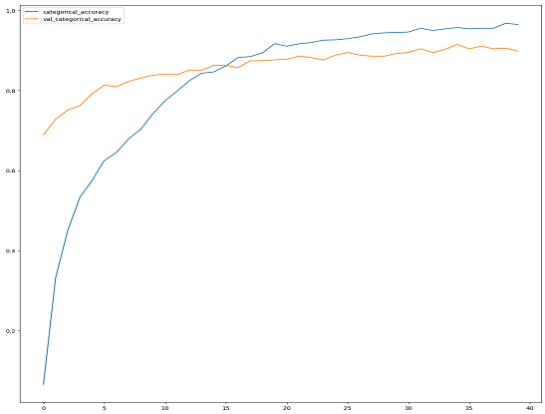
Accuracy

In the best case, for a deep learning model, we would like both curves to increase harmoniously during the training process, indicating that the model is learning and improving its performance on both the training and validation datasets. In Figure 3.11a training and learning curves intersect at around epoch 30 while in Figure 3.11b the intersection is earlier at around epoch

13. If the change continues to increase with validation and training accuracy diverging, this will signal that the model is overfitting. In the current case, while there seems to be a discrepancy, we do not believe that the model is overfitting. However, this shows that there is room to further investigate and improve performance.



(a) Accuracy of the model with an ensemble of EfficientNet and SwinT, trained for 40 epochs

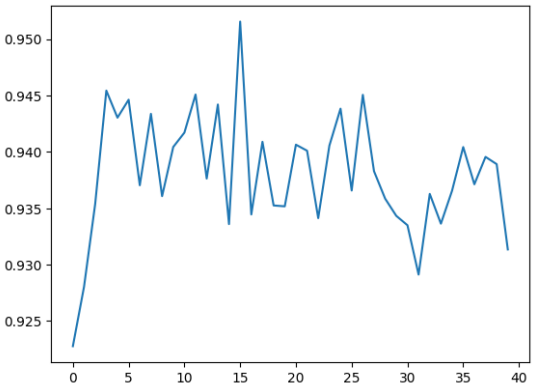


(b) Accuracy of the model with an ensemble of EfficientNet and BEiT, trained for 40 epochs

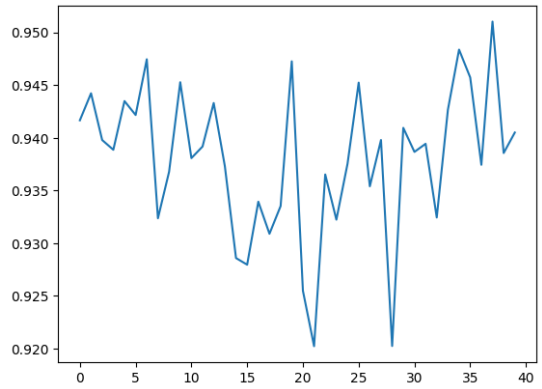
Figure 3.11: Accuracy curves of the two models being experimented

Macro F1-score

In Figure 3.12 we show how the Macro F1-score varies during the 40 epochs. Figure 3.12a shows that a high Macro F1-score is obtained very early, at epoch 15. However, a look at Figure 3.12b shows a peak at epoch 17 but it has another peak at epoch 38. This indicated that it is a good idea to investigate both models on unseen test data to have a better indication of which is most suited for the DFU disease classification.



(a) Macro F1 Score variation over the 40 epochs of training for the model with an ensemble of EfficientNet and SwinT, exhibiting a peak at epoch 15.



(b) Macro F1-score variation over the 40 epochs of training for the model with an ensemble of EfficientNet and BEiT, exhibiting a peak at epoch 38

Figure 3.12: Macro F1-score variation of the two models being experimented over 40 epochs

Summary of Metric and Comparison

In this section, a summary of the two models is shown in Table 3.4. For this table it can be seen that based on the main metric on which we are evaluating our model the EfficientNet and SwinT model has a higher macro F1-score compared to the EfficientNet and BEiT model. The class F1-score for none, infection is better for EfficientNet and BEiT while for ischaemia EfficientNet and SwinT is better. For both classes they have same class F1-score.

Table 3.4: Comparison of CNN and ViT siamese models

Model	EfficientNet and SwinT	EfficientNet and BeIT
Macro F1-score	**0.9516**	0.9510
loss	5.3856	**0.6980**
categorical_ accuracy	0.6922	**0.9556**
val_loss	2.9078	**1.9856**
val_categorical _accuracy	0.8270	**0.9110**
Accuracy	0.9320	**0.9395**
None F1-score	0.9265	**0.9370**
Infection F1-score	0.9217	**0.9308**
Ischaemia F1-score	**0.9783**	0.9565
Both F1-score	0.9798	0.9798
Macro Precision	**0.9477**	0.9472
Macro Recall	**0.9558**	0.9551
Macro AUC	0.9748	**0.9781**
Weighted Avg. Precision	0.9322	**0.9397**
Weighted Avg. Recall	0.9320	**0.9395**
Micro F1-score	0.9320	**0.9395**
Epoch #	**15**	38

The two models were evaluated on test data provided by the DFU2021 Challenge. This consists of 5734 unlabeled images that are used to make predictions and uploaded on the platform to get the required metrics.
Table 3.5 presents the performance of the two models on test data which were uploaded on the DFU2021 live challenge board. From the table, it can clearly be observed that the EfficientNet and BEiT model exhibits better overall performance in almost all the metrics except for the weighted average precision. Hence, the EfficientNet and BEiT model was further optimised. The predictions that showed the highest macro F1-score were averaged and loaded on the classification liveboard.

Table 3.5: Performance on Test data

Metrics	EfficientNet/SWINT	EfficientNet/BEiT
Macro F1-score	0.5850	**0.6160**
None F1-score	0.7442	**0.7478**
Infection F1-score	0.6072	**0.6149**
Ischaemia F1-score	0.5367	**0.5613**
Both F1-score	0.4520	**0.5401**
Macro Precision	0.5892	**0.6115**
Macro Recall	0.6368	**0.6570**
Macro AUC	0.8043	**0.8298**
Weighted Avg. Precision	**0.6818**	0.6728
Weighted Avg. Recall	0.6610	**0.6918**
epochs	16	**7**

When compared to the performance of related works, DFU-SIAM which is a model based on a siamese neural network for DFU disease classification, exhibits the best Macro F1-score as shown in Table 3.6. Galdran *et al.* [149] were actually the winners of the DFU challenge.

Table 3.6: DFU-SIAM comparison with related work

Metrics	DFU-SIAM	Galdran *et al.* [149]	Bloch *et al.* [156]	Ahmed *et al.* [169]	Qayyum *et al.* [168]
Rank	**BEST**	1st	2nd	3rd	4th
Macro F1-score	**0.6228**	0.6216	0.6077	0.5959	0.5691
None F1-score	0.7553	**0.7574**	0.7453	0.7157	0.7466
Infection F1-score	0.6276	0.6388	0.5917	**0.6714**	0.6281
Ischaemia F1-score	0.5495	0.5282	**0.558**	0.4574	0.467
Both F1-score	**0.5588**	0.5619	0.5359	0.539	0.4347
Macro Precision	0.5486	0.614	**0.6207**	0.5984	0.5814
Macro Recall	**0.6554**	0.6522	0.6246	0.5979	0.6104
Macro AUC	0.8599	**0.8855**	0.8616	0.8644	0.8488
W.Avg. Precision	0.6983	**0.7009**	0.6853	0.6730	0.68
W.Avg Recall	0.6815	**0.6856**	0.6657	0.6711	0.6636
Micro F1-score	0.6749	**0.6801**	0.6532	0.6714	0.6577
epochs	Avg(best epoch)	NA	NA	NA	5(ended)

3.2.5 Discussion

DFU classification is implemented using a Siamese Neural Network which is in itself a novel architecture, combined with Large Margin Cotangent Loss (LMCot) as a novel approach for enhancing performance in verification and identification. We further introduce the KNN classifier while iteratively searching for the best K while doing prediction on test data. These

are the reasons that explain why our model, DFU-SIAM, performs better than the other model in the related work. While Galdran *et al.* [149] focused on comparing Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) and achieved the best macro F1-score, our approach takes a different direction by combining these two architectures. By incorporating the strengths of both CNNs and ViTs, we capitalize on their complementary features and achieve improved results. As far as Bloch *et al.* [156] they used an ensemble of EfficientNet families with pseudo-labeling. In DFU-SIAM we choose EfficientNet, or more precisely, EfficientNetV2S, which is one of the best performing pre-trained CNN. Qayyum *et al.* [168] concentrated essentially on vision transformers. They propose the combination of two different pre-trained ViT models for feature extraction. For our proposed model, we chose BEiT, which is one of the best performing pre-trained transformers. However, we decided to make the last 10 layers of the BEiT transformer trainable as our experiments showed a significant increase in performance. DFU classification in our study used a novel approach of using innovative SNN architecture for classification of DFU. To further enhance its performance, we chose to use a novel approach called the Large Margin Cotangent Loss (LMCot) proposed by Duoung *et al.* [173]. Our proposed model, DFI-SIAM, surpasses the performance of other models discussed in related work.
Bloch *et al.* [156] employed an ensemble of EfficientNet models with pseudo-labeling, which differs from our methodology. Instead, we specifically chose the EfficientNetV2S model, known for its outstanding performance as a pre-trained CNN. We acknowledge, however, that the pseudo-labeling can be used in our model to further improve its performance.
Furthermore, Qayyum *et al.* [168] concentrated on Vision Transformers, proposing the combination of two distinct pre-trained ViT models for feature extraction. In our study, we adopt the BEiT model, which exhibits very good performance as a pre-trained transformer. However, we make a deliberate choice to train only the last 10 layers of the BEiT transformer to strike a balance between fine-tuning and computational efficiency.
By integrating these advancements and tailoring them to the specific requirements of DFU classification, the DFU-SIAM model achieves remarkable accuracy and sets a new benchmark in the field. It should be noted that the model's computational efficiency was not evaluated at this stage. This parameter is important if the model is to be deployed on ubiquitous devices. One limiting factor of the system is the imbalanced data, and this is also acknowledged by other researchers, with Bloch *et al.* [156] using pseudo-labeling and Generative Adversarial Network to tackle this.
While exploring need for more data, we may have clinics or medical centres that adhere to the idea of using deep learning models but are not willing to share the data with third parties. Ensuring patient privacy while integrating diverse datasets into a model has emerged as a significant limitation in deep learning research [180]. This problem can be addressed by deploying the model using Federated Learning [181]. The notable aspect of Federated Learning lies in its ability to handle data in a decentralized manner, thereby fostering a privacy-preserving environment in AI applications [182] in the event we require several distant sites

to contribute to having even more data, which is an important aspect for the training and implementation of a deep learning model.

3.2.6 Limitations

One limitation of our study arises from the substantial class imbalance present within the dataset, particularly evident in the under representation of "both" and the Ischaemia class. Upon careful inspection of the images, we observed that certain geometric data augmentation techniques were already applied to these classes during the dataset creation process. This imbalance has influenced the overall performance of the models. Nevertheless, it is worth noting that the DFU2021 dataset is currently the most comprehensive resource available for conducting research in this domain. Furthermore, we remain optimistic that with adequate computational resources, there is potential to explore additional variations and employ ensemble modeling techniques to enhance the outcome of our study.

3.2.7 Conclusion and Future Works

In this study, we have trained and tested a new model based on an ensemble of EfficientNet and BeIT Transformer in a SNN model that has outperformed some of the best results obtained for classification of DFU as detailed in related work 3.2.2. The dataset limitations can be addressed in future work by investigating the use of GAN which is a type of deep neural network that consists of two components: a generator network and a discriminator network [158]. Another option would be using pseudo labeling which is a technique used in machine learning to improve model performance by using unlabeled data in conjunction with labeled data [157]. The 5734 unlabeled data in the test image can thus be exploited.
This research marks an important step towards tackling the use of machine learning in the field of DFU image classification. Despite our limited processing power, we effectively utilized available resources to achieve significant results. With access to greater computational capabilities, we anticipate that further fine-tuning of our model will lead to even better performance.
As previously specified, there is a need to have a better-quality and more balanced dataset to curb data bias and ensure the model generalises well to unseen data. One possible solution that should be explored is accessing data collected at different geographically located medical facilities. This clearly poses the problem of data privacy, as the owner of the data would not want highly sensitive health-related data to be transferred to a third party. To overcome this barrier, the use of centralized Federated Learning or Peer-to-peer Federated Learning should be explored. Federated Learning, an innovative distributed interactive AI concept, holds exceptional promise in the realm of intelligent healthcare. This approach enables multiple clients, including entities like hospitals, to engage in AI training while upholding stringent data privacy protocols [183], [184], [185]. It entails the training of machine learning models across

datasets dispersed throughout various data centers, such as hospitals and clinical research labs, all while safeguarding data integrity [186].
Incorporating data from a variety of sources will undoubtedly contribute to enhancing the dataset's imbalance, thereby alleviating the data bias observed in the "both" and "ischaemia" classes. These classes currently exhibit only 621 and 227 occurrences, in contrast to the "infection" and "none" classes which encompass 2555 and 2552 instances, respectively. It is important to highlight that "both" and "ischaemia", which are the most serious forms of DFU, are relatively less prevalent in the samples. However, this poses a challenge for machine learning algorithms. One potential approach to addressing this imbalance is to employ GAN [122] for generating synthetic images. This technique has been successfully employed by Kim *et al.* [187] to augment liver ultrasonic image data using a semi-supervised approach.
This work serves as a stepping stone for future research and development aimed at effectively detecting, treating, and managing diabetic foot ulcers. Our ultimate goal is to contribute to advancements in the medical field, leading to improved patient outcomes and healthcare management. As the machine learning model learns by trying to reduce the loss to a minimum, it is prone to making erroneous predictions. If data bias is present, then there will most certainly be errors in predictions. Hence, from a medical point-of-view, it is mandatory to explainability on top of clinical validation [188]. The critical obstacle to the widespread acceptance of machine learning in healthcare and research relates to the black box nature of machine learning algorithms for the end user [189]. There is presently extensive research concentrating on Explainable AI (XAI), which aims to provide a suite of machine learning techniques that enable human users to understand, appropriately trust, and produce more explainable models [96]. This has to be given priority in any future work. One simple step could be to show a class activation mapping (CAM) approach that highlights the infected section or section with ischaemia and improves the visual interpretability [190]. Techniques like LIME (Local Interpretable Model-Agnostic Explanations) [191] and SHAP (Shapley Additive exPlanations) should be explored [192].
A crucial future direction, which we will present in the Section 3.3, for our research involves utilizing the Siamese Neural Network to develop a tool that can aid medical practitioners in evaluating the treatment protocols they administer to patients over time. This longitudinal disease evaluation tool would enable practitioners to monitor and adjust treatments as needed. Subsequently, after thorough testing and evaluation of the tool, it can be adapted into a preventive tool for early detection of DFU disease in patients, accessible via a mobile phone platform. In order to advance to the next phase, our plan involves collaborating with experts from public health research labs who possess the necessary expertise in designing protocols for assessing the effectiveness and acceptability of technology adoption in healthcare settings.

3.3 DFU-HELPER: Longitudinal DFU evaluation

3.3.1 Motivation

When addressing the prevention and treatment of diabetic foot ulcers (DFUs), the adoption of a multidisciplinary approach becomes crucial [193]. In their study, Netten *et al.* [194] investigated the reliability of utilizing mobile images for remote DFU assessment. Unfortunately, their experiment did not yield conclusive results, leading them to caution against relying solely on mobile images for making treatment decisions due to their low validity and reliability. They recommended that clinicians gather as much additional information as possible when utilizing such images. In contrast, the system we propose maintains the importance of clinical expertise while leveraging the power of deep learning to provide supplementary information for evaluating medical images. This additional information aims to enhance the clinician's ability to design personalized treatment protocols for individual patients. By combining the strengths of deep learning and clinical expertise, our system seeks to bridge the gap and provide valuable support to clinicians in making informed decisions regarding DFU treatment.
Recognizing the immense potential of applying deep learning to digital images, a collaborative venture involving Manchester Metropolitan University, Lancashire Teaching Hospitals, and the Manchester University NHS Foundation Trust has established an international repository comprising approximately 11000 diabetic foot ulcer (DFU) images. The primary objective of this repository is to facilitate the development of more advanced methods for DFU analysis. In our research, we utilize this dataset to train our algorithm and validate our model. The DFU challenge2021 organizers agreed to give access to this dataset [195]. Additionally, we incorporated datasets obtained from Kaggle [196], [197], [198] to further enhance our training process. This research aims at evaluating the progress of DFU disease using images. When comparing two-element vectors, there exist various alternative similarity techniques that can be employed, such as Euclidean distance, Pearson correlation coefficient [199], Spearman's rank correlation coefficient, and others [125]. The choice of technique for comparison depends on the specific objective we want to achieve. However, traditional similarity measurements may not be effective when dealing with complex datasets that exhibit diverse dimensions and characteristics and potentially require compression prior to processing. In such cases, Siamese Neural Networks (SNN) emerge as a promising solution. The architecture of Siamese Neural Networks was initially introduced in the early 1990s to address the challenge of signature verification as an image matching problem [124]. Figure 3.13 provides an initial overview of the operation of Siamese Neural Networks (SNN).
The structure of a Siamese Neural Network (SNN) consists of two identical artificial neural networks, A and A', as depicted in Figure 3.13. These networks are designed to learn the underlying representations of input vectors. Operating as feedforward perceptrons, they utilize error back-propagation during the training process to optimize their performance. Working in

parallel, the two networks generate outputs that are compared at the end using metrics such as cosine distance or Euclidean distance. The output of an SNN operation can be interpreted as the semantic similarity between the projected representations of the two input vectors.

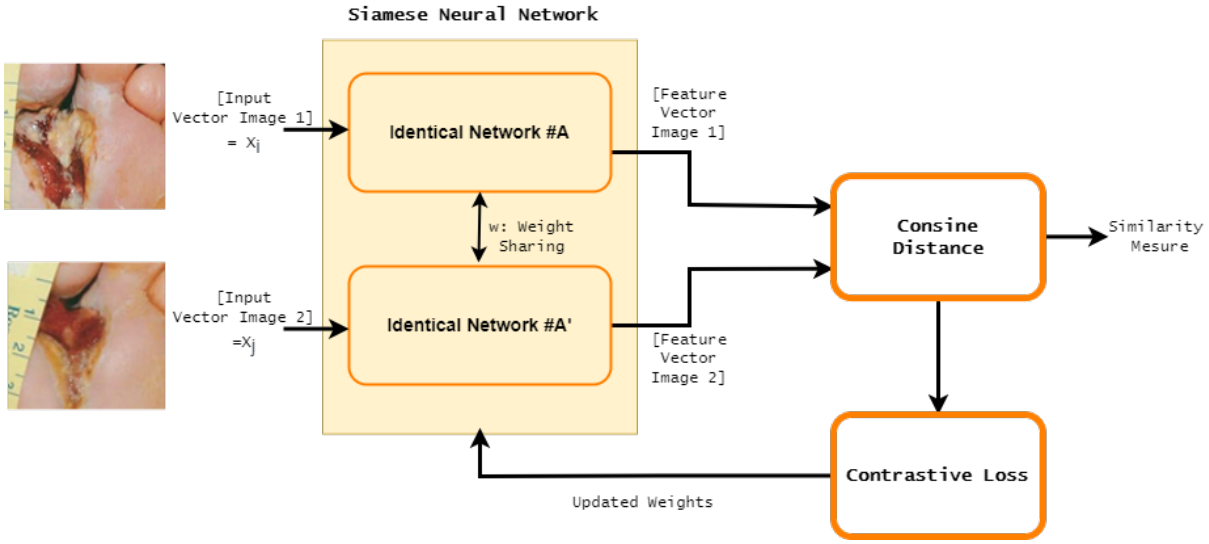


Figure 3.13: Block Diagram giving an overview of components of a Siamese Neural Network (SNN).

The contrastive loss [200] function is a type of distance-based loss function that updates the weights of a neural network in such a way that similar feature vectors have a minimal Euclidean distance. On the other hand, the distance is maximized between two dissimilar vectors. By employing the contrastive loss function, the neural network is encouraged to effectively separate and discriminate between different classes or categories based on their feature representations. More formally, we suppose that we have a pair of DFU images (I_i, I_j) and a label Y that is equal to 0 if the samples are similar and 1 otherwise. To extract a low-dimensional representation of each sample, we use a Neural Network which can be a CNN or Any ensemble model that encodes the input images I_i and I_j into an embedding space where $x_i = f(I_i)$ and $x_j = f(I_j)$. The contrastive loss is defined as:

$$L = (1 - Y) * || x_i - x_j ||^2 + Y * max(0, m - || x_i - x_j ||^2) \tag{3.1}$$

As mentioned in Section 3.3.2, our review of the existing literature reveals a noticeable scarcity of research focused on the application of deep learning specifically for assessing the progression of Diabetic Foot Ulcers (DFUs) over time. While several classification techniques have been investigated, there is a notable gap when it comes to considering the temporal dimension of DFU evaluation. This research aims to address this critical gap by exploring the potential of Siamese Neural Networks for tracking and evaluating the development of DFUs over time. By filling this research void, we aim to provide valuable insights into the effective utilization of deep learning in the longitudinal assessment of DFUs.

This research investigates the application of Siamese Neural Networks (SNN) for the longitudinal follow-up of patients with Diabetic Foot Ulcers (DFUs) who have undergone a treatment protocol under the guidance of a clinician. The DFU-Helper Framework helps clinicians gain better insight into the progression of DFU diseases and take corrective measures. This section makes the following key contributions:

1. A novel Siamese Neural Network model validated in terms of performance against other models in terms of performance metrics.
2. Introduction of a valuable tool for clinicians to validate the efficacy of treatment protocols used for DFUs by harnessing the similarity learning capabilities of the Siamese Neural Network.

Overall, this research contributes to the field by presenting a novel approach using Siamese Neural Networks for assessing and validating treatment protocols for DFUs. This has the potential to enhance the personalized management of DFU patients and improve their overall outcomes. It also answers the research question:RQ-1: How can deep learning be applied for DFU management?.

3.3.2 Related Work

This section investigates primarily the application of Siamese Neural Networks (SNNs) in the context of utilizing medical images for longitudinal disease evaluation. As a secondary objective, we extend our focus on the use of SNNs within the medical domain with the aim of gaining insights into their architectural adaptations that are more suited to our specific research objectives.
Li *et al.* [201] used a Siamese Neural Network to monitor and assess the severity and progression of medical imaging in two specific diseases: retinopathy of prematurity (ROP) in retinal photographs and osteoarthritis in knee radiographs. The technique employed measures the similarity between two images captured at different time points. As no severity ranking labels were available, the authors computed a median Euclidean distance from a set of known normal images. Their approach involved using a convolutional Siamese network with a ResNet-101 [202] architecture that had been pre-trained on the ImageNet dataset. To accommodate the specific requirements of their algorithms for the retina and knee, the final fully connected layer of ResNet-101 was modified to output three or five nodes, respectively, from each sub-network. The implementation of Li et al. was conducted in Python, utilizing the Adam optimizer with a learning rate of 5×10^{-7}. During training and validation, a batch size of 16 was employed, and the model was saved based on the lowest validation loss for subsequent testing evaluation. The results obtained from their experiments indicate that the utilization of Siamese Neural Networks and Euclidean distance measurements enables the representation of disease severity on a more nuanced and continuous spectrum compared to traditional categorical disease classification systems.
In a recent study by AbdulRaheem *et al.* [203], the authors propose the use of Siamese Neural Networks (SNNs) for the continuous evaluation of eye disease severity. Similar to the work of

Li *et al.* [201] mentioned earlier, they employ a twin-CNN architecture. The proposed system is specifically demonstrated in the domain of diabetic retinopathy. However, it is important to note that AbdulRaheem et al. utilize a Siamese Triplet network, which aims to determine the distance between image embeddings. For the sub-network implementation, ResNet-101, a convolutional neural network architecture, was used. They incorporate a triplet mining algorithm, where the data are provided as triplet image pairs comprising an anchor image, a positive image, and a negative image. Each image pair is passed through the pre-trained network, which learns the distributed embedding of the images based on their similarities and dissimilarities with respect to the anchor image. The Euclidean distance between the images is then computed from the final connected layer, representing the difference between the images. This distance serves as an abstraction of the severity score for the respective image.
In their study, Akbar *et al.* [204] proposed the use of Siamese Convolutional Neural Networks (CNNs) for the assessment of the continuous spectrum of lung endema severity using chest radiographs. Unlike the previous works mentioned, they employed a pre-trained CNN architecture called DenseNet121 [102] instead of ResNet-101 [202]. The authors utilized the Euclidean distance as a measure of similarity between images. For the optimization process, the Adam optimizer was chosen for all models, with a learning rate of 2e-5. Model weights were saved at each epoch if the validation loss decreased. If the validation loss plateaued or did not improve for more than 10 epochs, early stopping was applied. The results showed that their model successfully assessed the severity of pulmonary edema from chest radiography. To label the dataset used in this research, two certified radiologists participated in the study. The authors investigated the performance of their model using four different loss functions: contrastive loss, mean square error (MSE) loss, Huber loss, and a combination of contrastive and MSE loss.
Fiaidhi *et al.* [205] leverage the characteristics of SNNs needing small samples of data for training. They introduced a Siamese neural model that uses a triplet loss function that enables the gastroenterologist to inject anchor images that can correctly identify the ulcerative colitis severity classes. To monitor the severity over time, a triplet loss function is applied. Gastroenterologists inject anchor images that can correctly identify the ulcerative colitis severity by using the classes and using the Mayo Clinic Ulcerative Colitis Endoscopic Scoring scale.
During our research, we also explored non-deep-learning-based methods for assessing image similarity in the context of medical images. Hu *et al.* [206] proposed a method that relies on feature extraction and analysis. They applied their method to rat brain histological images and compared the similarity estimates with expert evaluations to demonstrate its effectiveness. Their approach involved various computer vision techniques, such as color model conversion, image normalization, anti-noise filtering, contour detection, conversion, and feature analysis. The feature search process utilized an anchor image. Inonescu *et al.* [207] conducted a study on image similarity using 54 video files of endocapsules labeled by gastroenterologists. They employed techniques based on color histogram and Local Binary Patterns (LBP) Histogram and calculated the difference between image pairs to determine similarity. A value close to

0 indicated higher similarity. These studies demonstrate alternative approaches to assessing image similarity in medical imaging, utilizing feature-based methods and computer vision techniques instead of deep learning-based approaches. Now, we move on to explore the use of SNNs in the medical domain applied to images.

These studies presented above show methodologies for evaluating image similarity in medical imaging, employing feature-based methods and computer vision techniques as alternatives to deep learning-based approaches. Now, we move our focus to investigating the application of Siamese Neural Networks (SNNs) in the medical domain, specifically in the medical field, using image analysis.

Ornob *et al.* [208] introduced a Siamese few-shot learning model for early detection of COVID-19, aiming to mitigate the long-term effects of this dangerous disease. Their proposed architecture combined few-shot learning with an ensemble of pre-trained Convolutional Neural Networks, enabling the extraction of feature vectors from CT scan images for similarity learning. They implemented a Triplet Siamese Network for classification, utilizing six transfer-learning-based models (ResNetV2, DenseNet, SwinTransformer, MobileNetV2, EfficientNetB0, ResNeXt-101) as the backbone of the network to create an ensemble model. This ensemble model generated embeddings for each image in the input triplet, enhancing the accuracy of the classification. Mehboob *et al.* [209] implemented a Siamese Neural Network with a VGG-16 backbone and CNN for multiclass classification of Alzheimer's Disease. They utilized the network to classify different stages of the disease. Zeng *et al.* [210] proposed a novel Siamese Convolutional Neural Network (CNN) architecture using InceptionV3 [211] as the backbone. Their work focused on binary classification, and they employed the Adam optimizer for training. Vasconcellos *et al.* [212] conducted research on the classification of heartbeats using 12-Lead ECG datasets. They developed a Siamese Neural Network based on CNN for this task. Table 3.1 provides a summary of other related works in the medical field that have explored the use of Siamese Networks.

Table 3.7: Summary of use of Siamese Neural Network in Medical Field

Research work	Year	Field of Application	SNN Architecture	Objective	Data
Wang *et al.* [213]	2017	Cancer: Spinal metastasis	3 sub-network of 5 layers of CNN	Detection	MRI images
Hajamohideen *et al.* [214]	2023	Alzheimer's disease	Triplet Loss Function / CNN with VGG16 backbone / K-NN for neighboring analysis	Multi-Class Classification	MRI images
Tummala *et al.* [215]	2023	Blood Cell	EfficientNet Backbone / Contrastive Loss	Binary Classification	Microscopic images
Shorfuzzaman *et al.* [216]	2021	Covid19	Pretrained CNN VGG-16 /Constrastive Loss	Binary Classification	CT-Scan
Ahuja *et al.* [217]	2022	Covid19	Pretrained CNN ResNet18 /Contrastive Loss	Multi-Class Classification	CT-Scan
Cueva *et al.* [218]	2022	Osteoarthritis	ResNet-34 architecture modified with two fully connected (FC) layers added	Detection and Classification	X-Ray Images

Table 3.7 displays additional research where a Siamese Neural Networks is employed. The purpose of this investigation was to explore architectures utilized in the medical domain and assess their suitability for our requirements.
After working on the related work section, It is worth noting that no relevant research has been found in the literature regarding the evaluation of diabetic foot ulcers (DFUs) over time using deep learning techniques. While studies have addressed classification and detection techniques for DFUs, none have been found to engage SNN. In terms of longitudinal use of SNN we have also come across very little research work. We can conclude that the investigation of DFU evaluation over time using deep learning remains largely unexplored.

3.3.3 Proposed System

DFU-Helper Overall Framework

The overall framework, DFU-Helper, depicted in Figure 3.14, illustrates the proposed approach for monitoring the treatment progress of a diabetic foot ulcer patient, as initiated by a medical practitioner.

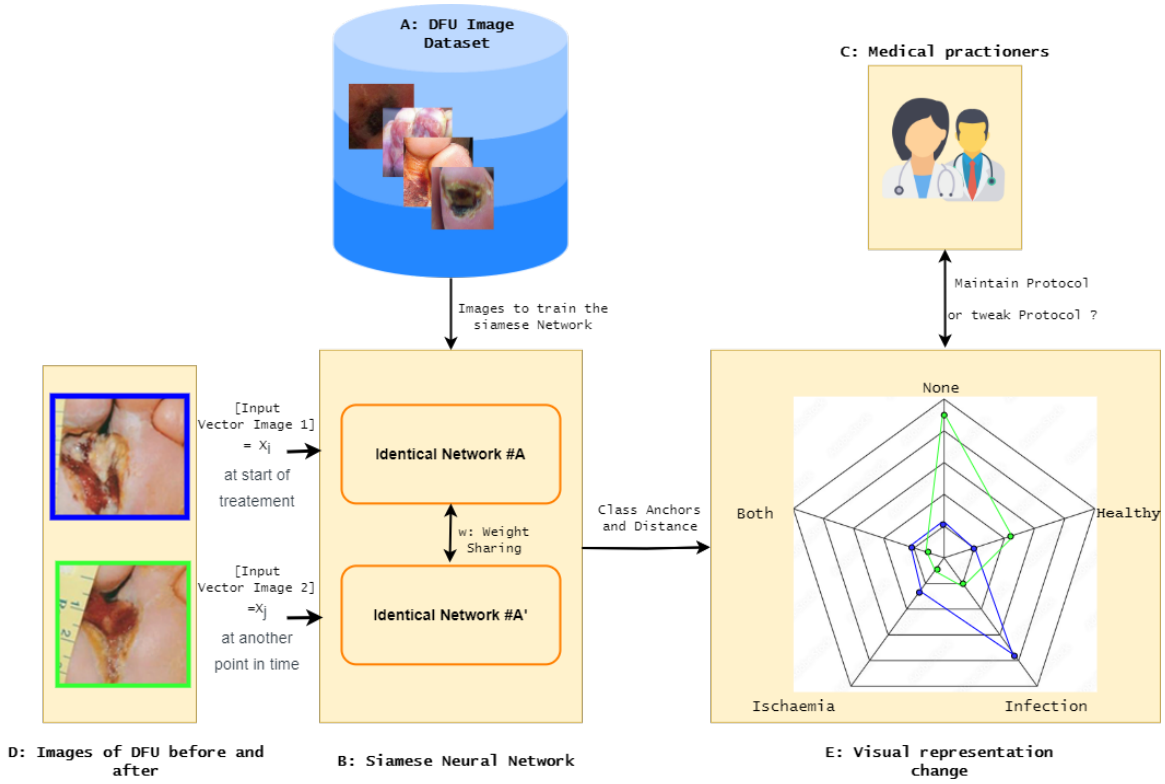


Figure 3.14: Overall schematics of the DFU-Helper framework for DFU longitudinal evaluation of DFU before and after the start of treatment by a medical practitioner.

The following gives an exhaustive description of the proposed DFU-Helper framework:

1. Obtain a representative dataset of diabetic foot ulcers (DFU) that provides a comprehensive description of different types of ulcers.
2. Train a Siamese Network using similarity learning on the dataset. This network will learn to distinguish between dissimilar images and group similar images together.
3. Test the SNN to validate its performance in terms of regrouping similar items closer and dissimilar items further. We plan to perform this step by testing it on a classification task and comparing it with known models.
4. Create anchor points for each significant class, including healthy, ischaemia, infection, both, and none.
5. When a medical practitioner examines a patient for the first time, they capture an image of the wound, which is represented by the first image in Figure 3.14. This image is marked with a blue border, indicating the pre-treatment stage. The practitioner then initiates an individualized treatment protocol.
6. During the subsequent visit, the doctor performs the standard evaluation procedure and captures another image of the wound. This new image is then fed into the system, along with the initial image, for further analysis and comparison.

7. The system generates a radar chart that plots the distances between the images and each of the anchor classes, as well as a table with the distances from the anchors.
8. In Figure 3.14, we can observe that the blue line representing the initial image shows a closer similarity to the infection class. In the second image, the blue line is plotted as being more similar to the healthy class, indicating an improvement and a movement towards the normal class.

Having introduced the Siamese Neural Network that implements similarity learning, the subsequent section will delve into an explanation of what exactly similarity learning entails.

Similarity Learning

The main goal of the learning process is to adjust the parameters in order to minimize the distance between encoded features of similar input image pairs while simultaneously maximizing the distance between dissimilar image pairs. The choice of loss function for training depends on the image pairs, their associated labels, and the specific parameterized distance function being used. Figure 3.15 provides an overview of the training process and feature vector extraction in the SNN. The anchor point for each class is computed as the average of the feature vectors belonging to that class. The sub-networks within our SNN consist of an ensemble of CNN backbones and Vision Image Transformers (ViT). Transformers have revolutionized NLP by addressing the limitations of sequential data tasks previously handled by RNNs. Transformers gained prominence through the influential paper "Attention Is All You Need", which leveraged self-attention mechanisms to capture contextual information in sentences [15].

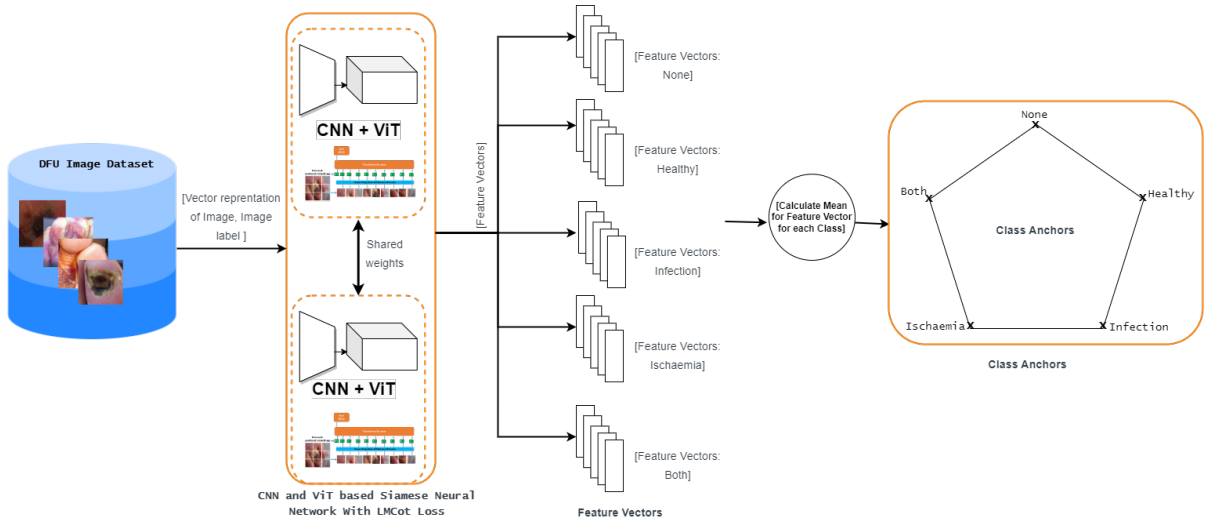


Figure 3.15: Similarity learning with CNN-ViT sub network and Large Margin Cotangent Loss (LMCot) [173].

The loss function implemented in the Siamese network is not the same as the loss function in a traditional artificial neural network. In the proposed SNN, we choose to use the Large Margin Cotangent Loss, which we will explain in the next section.

Loss function for Training the SNN

The loss function used during the training phase was the Large Margin Cotangent Loss (LMCoT) [173]. The primary motivation behind LMCot is to address the limitation of the cosine function used in existing methods such as ArcFace [219]. The cosine function returns values between $[-1, 1]$, which limits its ability to accurately reflect the angle between vectors. In contrast, the cotangent function has an unrestricted range of values, making it more suitable for measuring angles.
The LMCoT loss function is defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cot(\theta yi+m))}}{e^{s(\cot(\theta yi+m))}+\sum\limits_{j=1,j\neq yi}^{n}e^{s\cot\theta ji}}, \tag{3.2}$$

where L represents the LMCot loss, N is the number of samples, s is a scale parameter, m is the margin, θyi is the angle between the weight and feature vector of the ground truth class, and θji is the angle between the weight and feature vector of class j.
To calculate the cotangent values, the LMCot loss function utilizes the l_2-normalized feature vectors and weights. The loss function penalizes the difference between the cotangent of the ground truth angle θyi and the cotangent of the angles θji for other classes. This encourages the model to optimize the decision boundary to improve classification accuracy.
In order to facilitate the comparison of a new DFU image, a reference point is essential. This reference point is referred to as the "class anchor." The forthcoming section will explain the concept of class anchors and outline the methodology for their calculation.

Class Anchors

After training and validation of the SNN, the feature vectors of images belonging to each class are extracted and averaged, resulting in the generation of class anchors. During the testing phase, an input image is passed through one of the sub-networks of the SNN, which encodes it into a feature vector. In the lower-dimensional feature space, the feature vector of the test image is compared with the feature vectors of all the training samples using distance measures. This distance measure is utilized to plot a radar chart, providing a visual representation of the initial characteristics of the image. When a second image is inputted into the SNN, another

plot is generated on the same radar chart, visually indicating the progression of the disease.

$$Anchor\ classx = \frac{1}{n}\sum_{k=1}^{n} x_k = x_1 + x_2 + \cdots + x_{n-1} + x_n \tag{3.3}$$

With the class anchors established, the subsequent step involves calculating the similarity between a new DFU image and the class anchors. The next section elaborates on this process.

Similarity Function between Test Image and Anchors

To assess similarity, both cosine distance and Euclidean distance are employed between the feature vector of an image and the class anchors, as depicted in Figure 3.16. The cosine similarity function given by equation (3.5) is utilized, and the calculation of distance includes both cosine distance, equation (3.4) and Euclidean distance, equation (3.6). These measures highlight the disparity between the feature vector of the test image and the feature vector of the anchor class.

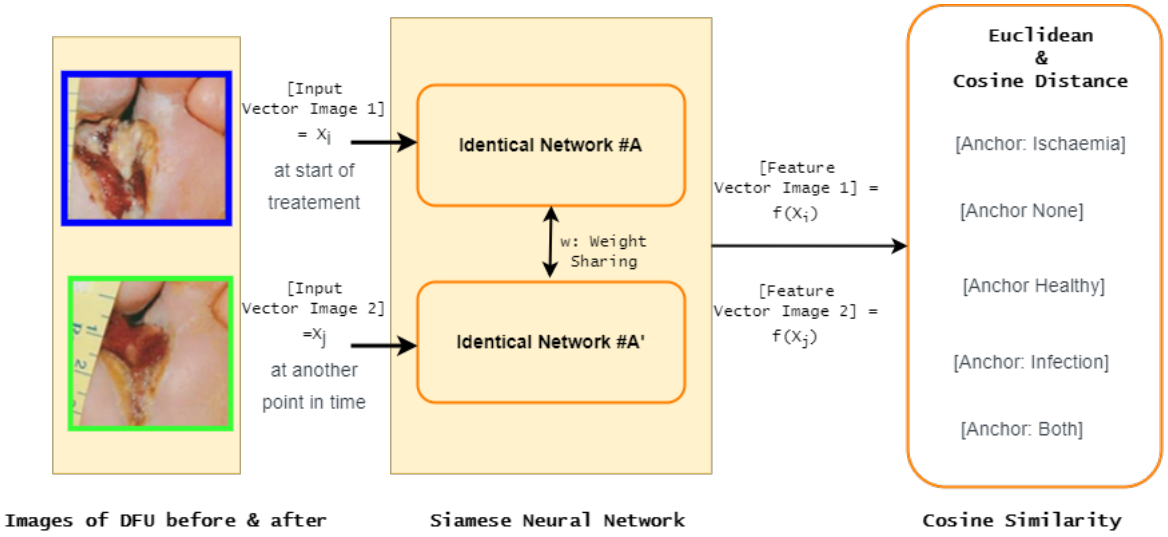


Figure 3.16: Block Diagram Summarizing how the similarity is calculated and plotted once the anchors of the classes are known and a test image is received.

$$Cosine\ Similarity(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \tag{3.4}$$

$$Cosine\ Distance(A, B) = 1 - CosineSimilarity(A, B) \tag{3.5}$$

$$Euclidean\ Distance\,(A, B) = \sqrt{\sum_{i=1}^{n} (B_i - A_i)^2} \tag{3.6}$$

Figure 3.17 gives a visual explanation of the cosine similarity and Euclidean distance. The cosine similarity represents the angle between two vectors, and the Euclidean distance represents the distance between two points in Euclidean space.

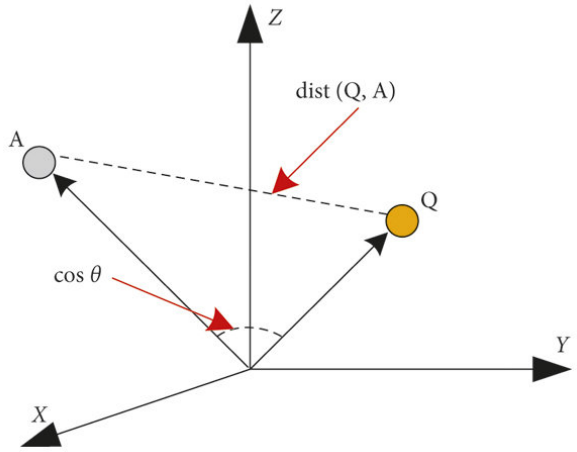


Figure 3.17: Visual illustration of Cosine Similarity and Euclidean Distance [220].

Now that the entire proposed process has been explained, it is imperative to dive into the rationale underlying the incorporation of class anchors within the framework.

Justification for using Class Anchors

When utilizing the SNN for disease evolution assessment, the conventional representation of similarity between two images, as depicted in Figure 3.13, does not suffice for our purpose. To illustrate this, we present a hypothetical scenario in Figure 3.18, Figure 3.19, and Figure 3.20, where we compare two images, Image A and Image B, belonging to the same patient at different time points. In this context, direct image comparison using a similarity score alone does not necessarily provide insights into whether the situation is improving or deteriorating. Hence, we demonstrate the necessity of comparison with an anchor image to discern the progress or regression of the disease

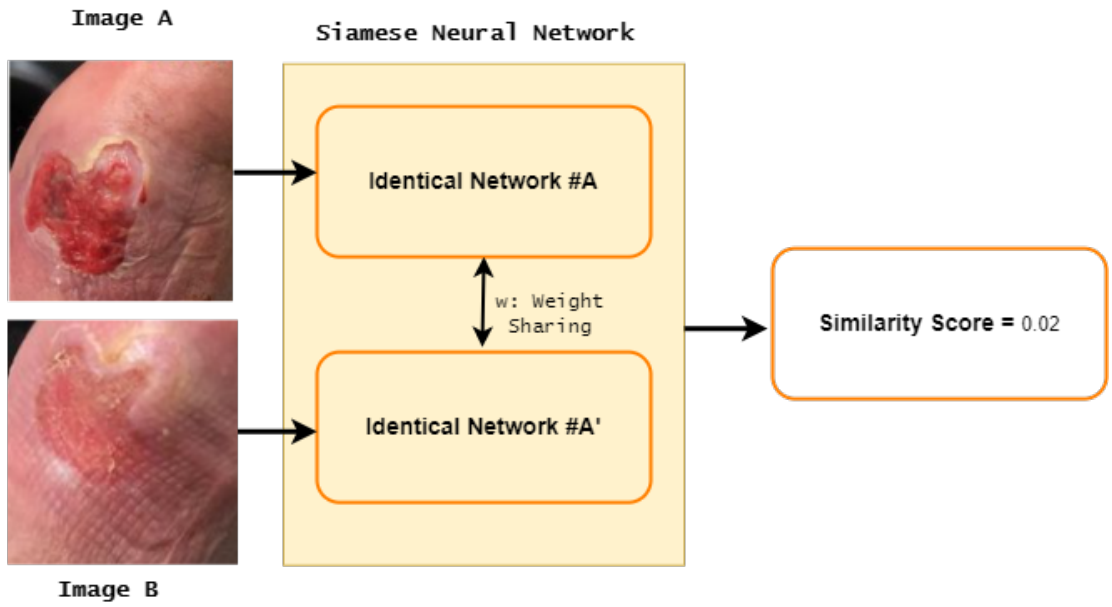


Figure 3.18: Example 1: When we input Image A and Image B into the SNN, the calculated similarity score is 0.02, indicating a low level of similarity between the two images. However, this score alone does not provide any indication regarding whether the condition is improving or worsening. This highlights the necessity of refining the model to gain better insights into the disease's progression.

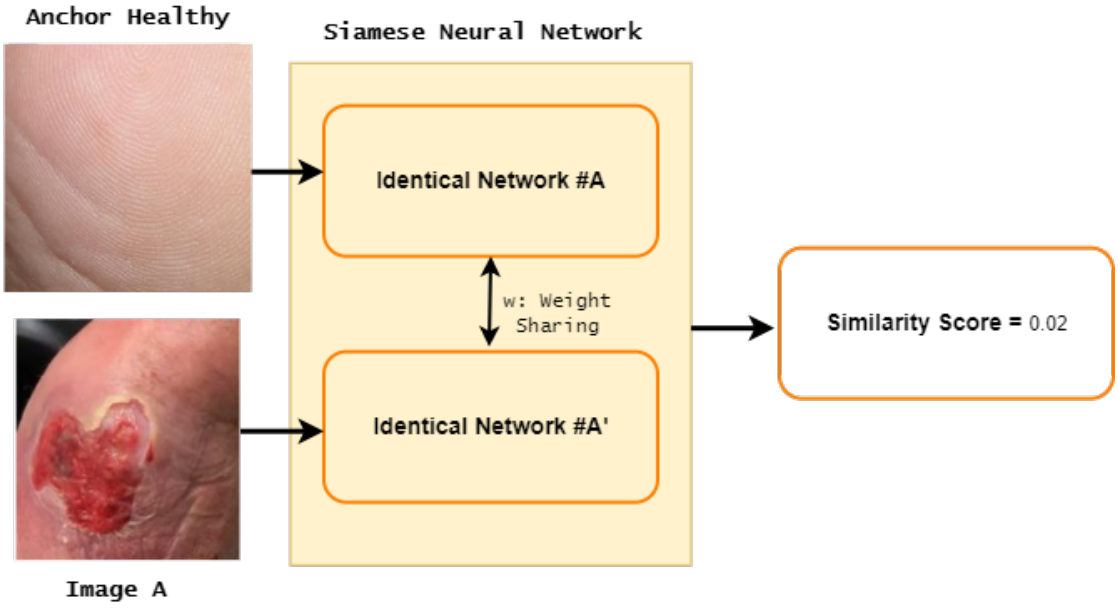


Figure 3.19: Example 2: We teak previous example, and give our SNN an Anchor representing Healthy image and Image A as a DFU image. We get a similarity score of 0.02, which means very low similarity. This means we have a situation that is clearly far from being healthy, and immediately the information becomes more mean.

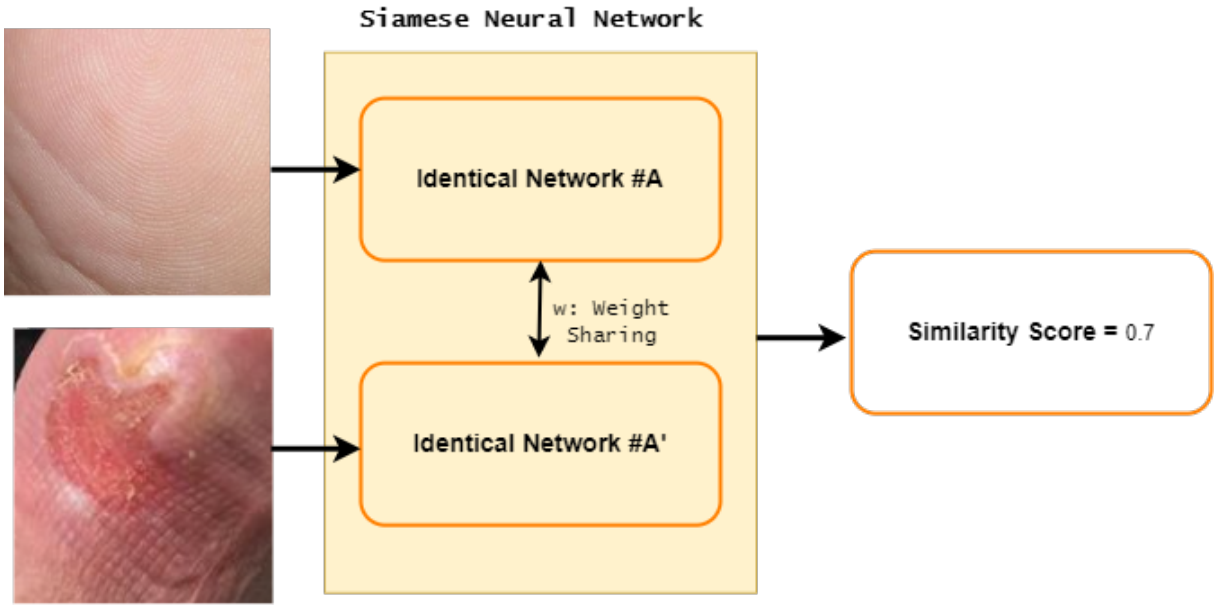


Figure 3.20: Example 3: Using the same Healthy Image as an anchor, we feed Image B into our SNN and obtain a similarity score of 0.7. This high similarity score indicates a significant improvement compared to the previous score of 0.02, with respect to the Healthy Anchor. This provides a clear insight that the treatment protocol is yielding the expected outcome, as the similarity between Image B and the Healthy Anchor has increased substantially.

The DFU-Helper framework is a novel system that utilizes a siamese neural network for implementing similarity learning. The SNN in DFU-Helper consists of subnetworks, which are ensembles of CNN and Vision transformers. The training of the SNN is performed using the LMCoT loss function, specifically designed for similarity learning. This framework represents a pioneering effort in the field of the application of deep learning for the longitudinal evaluation of DFU diseases over time.

3.3.4 Experimentation and Results

Dataset

Data quality is a crucial factor that directly affects the performance of supervised learning algorithms. The utilization of a representative and high-quality dataset is critical for achieving optimal accuracy and performance [175]. In this study, we obtained the dataset from the DFU2021 challenge organized by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) society [80]. The proper licensing was also secured for this research, ensuring that all ethical and legal requirements were met. We further added an addition class called healthy, which is available on Kaggle platform [196], [198], [197]. Upon initial preprocessing, we observed that the dataset's class distribution was imbalanced, with 621, 2555, 227,

and 2552 instances belonging to the classes both, infection, ischaemia, and none, respectively. The number of images of normal/healthy classes we obtained from Kaggle was 543. The reason to introduce the healthy class by collecting additional images from Kaggle is that the none class represents ulcers without any infection or ischaemia [80], therefore a healthy condition anchor was still required to assess ulcer evolution towards recovery. Sample images from the different classes are shown in Figure 3.21 and give a better idea. From the class distribution, we can conclude that the dataset is imbalanced. Such an imbalance poses a challenge to the performance of supervised learning algorithms, as they tend to be biased towards the majority class. To address imbalanced data, we applied geometric data augmentation techniques.
To test the proposed system, it was mandatory to have the proper hardware and software setup. The next section focuses on the experimental setup.

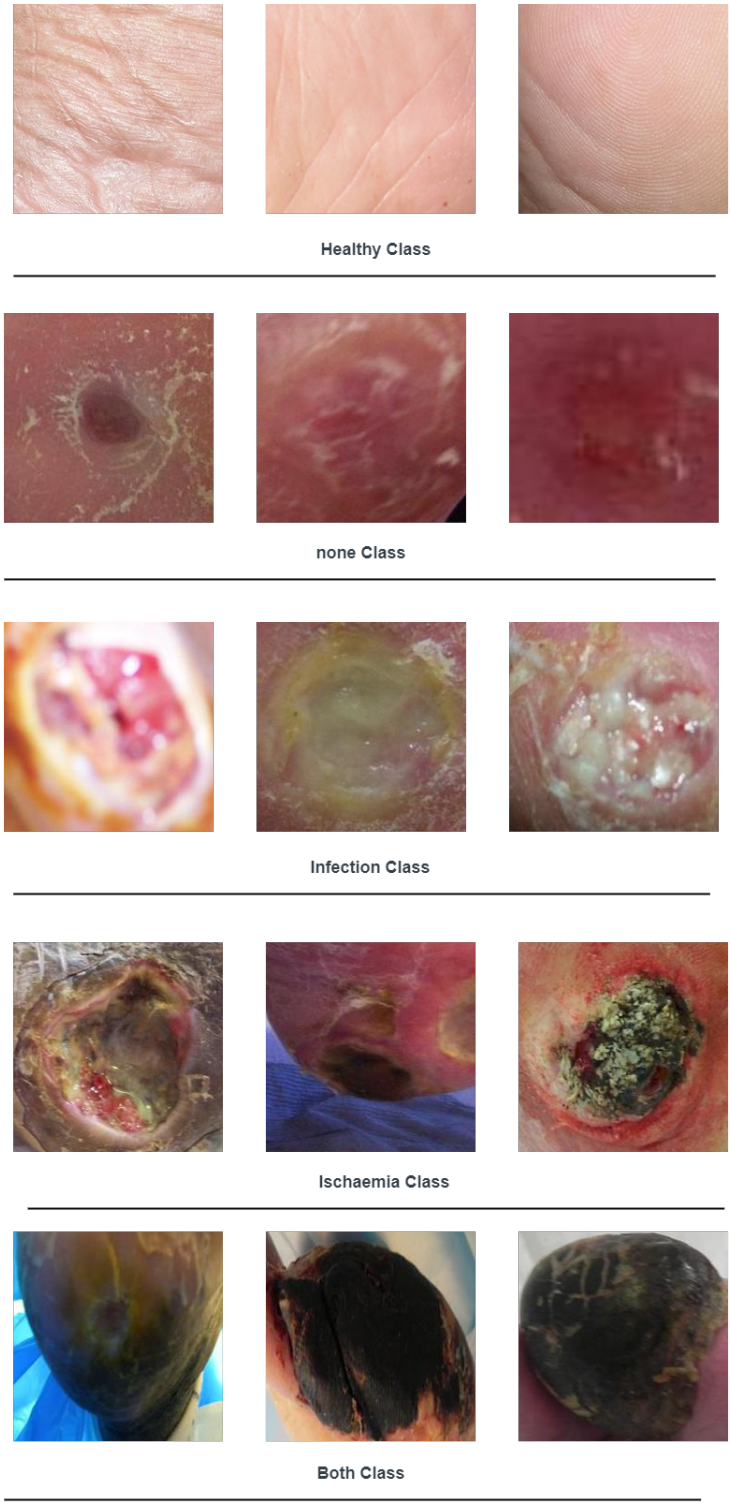


Figure 3.21: Example of images in each classes of DFU.

Experimental setup

The experiments were conducted on a Windows 10 Pro operating system, running on a powerful hardware configuration comprising 64 GB of RAM and an Intel(R) Xeon(R) W-2155 CPU operating at 3.30 GHz. The system was further enhanced with an NVIDIA GeForce RTX 3060 GPU, boasting 12 GB of dedicated memory. To facilitate the experiments, the system was configured with CUDA version 11.7, Tensorflow 2.10.0, and Python 3.10.9.
The selection of hyperparameters in this study was influenced by available computational resources. The batch size was set to 8, and the input images were resized to dimensions of 224 by 224 pixels with RGB channels. The model was trained for 40 epochs. A fixed learning rate of 10^{-6} was employed.
As explained in the previous section, the DFU-Helper framework uses a Siamese Neural Network, which is composed of two identical sub-networks. The following section explains the backbone used for the sub-networks.

Siamese sub-network backbone

For the CNN backbone, we use EfficientNetV2S based on EfficientNet [152] architectures, which have been shown to significantly outperform other networks in classification tasks while having fewer parameters. EfficientNetV2S has fewer parameters, making it more suitable for low-resource settings, and it uses a combination of efficient network design and compound scaling to achieve high accuracy with fewer parameters [171].
The second backbone of the ensemble model is based on Vision Transformers. This was first introduced by the paper "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale" [115], and is referred to as Vision Transformers (ViT). We chose Bidirectional Encoder representation for Image Transformers (BEiT), which uses a pre-training task called masked image modeling (MIM).
Having established all the parameters, the testing of DFU-Helper was conducted, and the next section provides a detailed explanation of the results obtained.

Results

In order to validate the performance of our model for monitoring image similarity against class anchors, we need to establish a method for evaluation. To achieve this, we decided to compare the classification capabilities of our model with published works on the same DFU dataset. After training our model and obtaining predictions on the test data, we applied a pseudo-labeling technique to the test data and retrained the model to further optimize its performance. Pseudo-labeling [157] was performed using a threshold of 0.9 to ensure a well-balanced model. This approach allowed us to assess the effectiveness of our model and refine its performance.

Comparison SNN in DFU-Helper framework

As discussed in previous sections, the DFU-Helper framework employs a Siamese Neural Network (SNN) for implementing similarity learning. It is crucial to assess the SNN's reliability when dealing with DFU images. Since there is no existing work specifically focusing on SNN for DFU, we conducted a classification test using our trained SNN for similarity learning on a dataset of 5734 test samples. The predictions were then uploaded to the online platform provided by the organizers of the DFU2021 challenge. The results, presented in Table 3.8, demonstrate the exceptional performance of our model across all evaluated metrics. This outcome signifies that our model effectively distinguishes between similar and dissimilar images.

Table 3.8: SNN used in the DFU Framework compared to published work in literature on classification task

Metrics	**Our SNN**	**Galdran *et al.* [149]**	**Bloch *et al.* [156]**	**Ahmed *et al.* [169]**	**Qayyum *et al.* [168]**
Rank	**BEST**	1st	2nd	3rd	4th
Macro F1-score	**0.6455**	0.6216	0.6077	0.5959	0.5691
None F1-score	**0.7607**	0.7574	0.7453	0.7157	0.7466
Infection F1-score	0.6348	0.6388	0.5917	**0.6714**	0.6281
Ischaemia F1-score	**0.6189**	0.5282	0.558	0.4574	0.467
Both F1-score	**0.6009**	0.5619	0.5359	0.539	0.4347
Macro Precision	**0.6507**	0.614	0.6207	0.5984	0.5814
Macro Recall	**0.6697**	0.6522	0.6246	0.5979	0.6104
W. Avg. Precision	**0.7136**	0.7009	0.6853	0.6730	0.6800
W. Avg. Recall	**0.6931**	0.6856	0.6657	0.6711	0.6636
W. Avg. F1-score	**0.6839**	0.6801	0.6532	0.6714	0.6577

In Table 3.8, we present the application of the Siamese Neural Network (SNN) within the DFU-Helper Framework for the classification task. This comparison encompasses an evaluation against leading research endeavors that have leveraged DFU images for reference. Our SNN, trained using similarity learning, demonstrates superior performance compared to the results achieved by other researchers, with the exception of the infection F1-score.

Results of DFU-Helper framework

To demonstrate the effectiveness of the experiments, we obtained images from several published articles [221], [222], [223] and we tried to compare if the findings of the model correlated with the descriptions given in the articles we are using as baseline. We present the results of multiple use cases using both cosine similarity and Euclidean distance calculations. By utilizing these different measures of similarity, we aim to provide a comprehensive analysis and ensure the reliability of our conclusions.

Use-Case #1

For the initial application case, we obtained images from the study conducted by Dayya *et al.* [221]. The serial images, depicting the measurements of a Wagner grade 2 wound with progressive healing in a diabetic patient (A-C), are presented in Figure 3.22.
The performance of our model was evaluated using cosine similarity and Euclidean distance metrics, as illustrated in Figures 3.23 and 3.24, respectively. Upon processing image A, the model identified it to be highly similar to infection, with a cosine similarity score of 0.81 and a semantic similarity score of 0.41 using Euclidean distance. In both cases, image C was deemed similar to the anchor class "none and healthy".
While the results seem obvious between images A and C, this is not the case for images A and B. While the radar plot gives a good visual and quick understanding of disease progression, we need to closely analyze the detailed results of both similarity metrics we are using.
From Table 3.9, we observe that image A, upon arrival, has a similarity score of 0.81 with the infection class, which aligns with the description of the wound. However, for image B, the similarity score shows less similarity with the infection class and higher similarity with both the healthy and ischaemia classes. After 12 weeks of treatment, Image C clearly exhibits similarity to neither infection nor ischaemia but rather to the healthy class. Similar trends can be seen in Table 3.10.

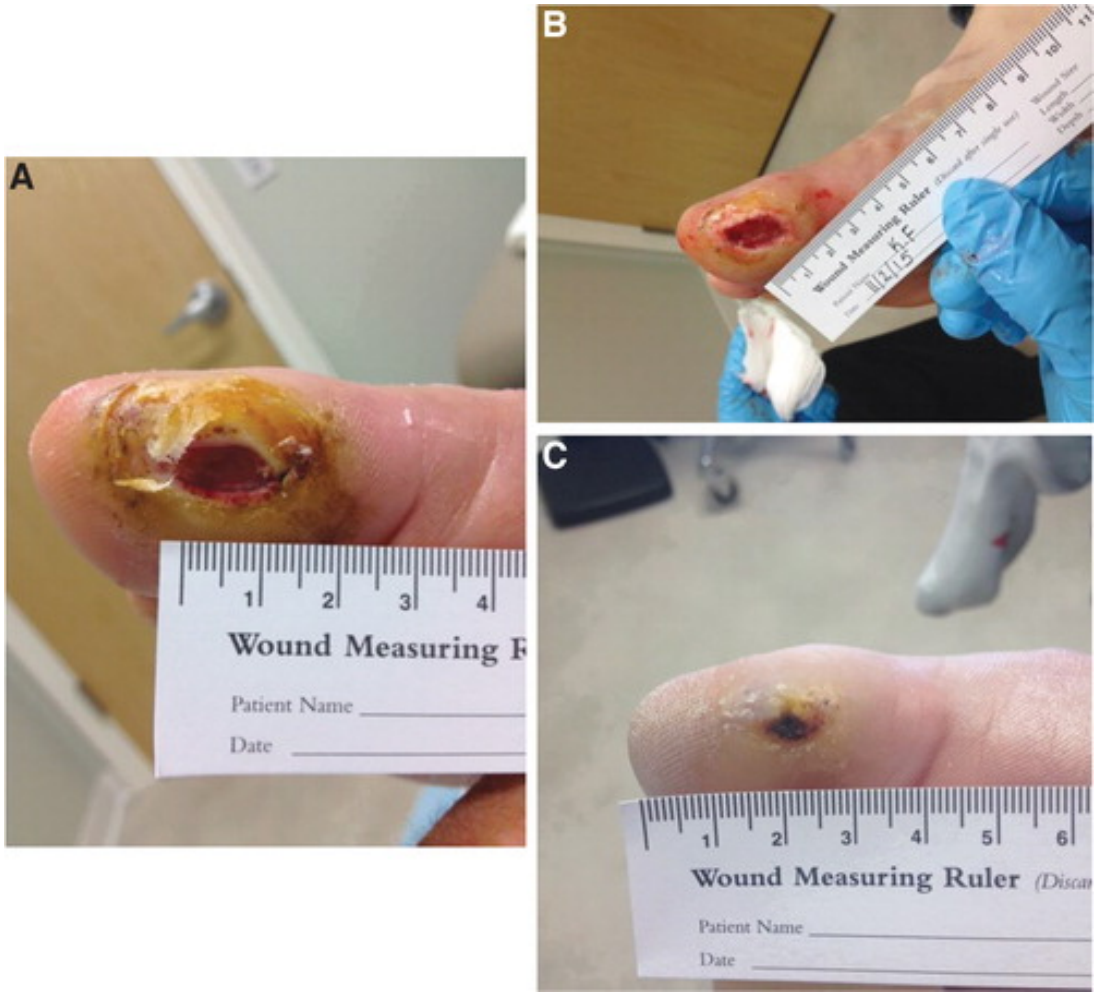


Figure 3.22: DFU images for Use-Case #1 [221].

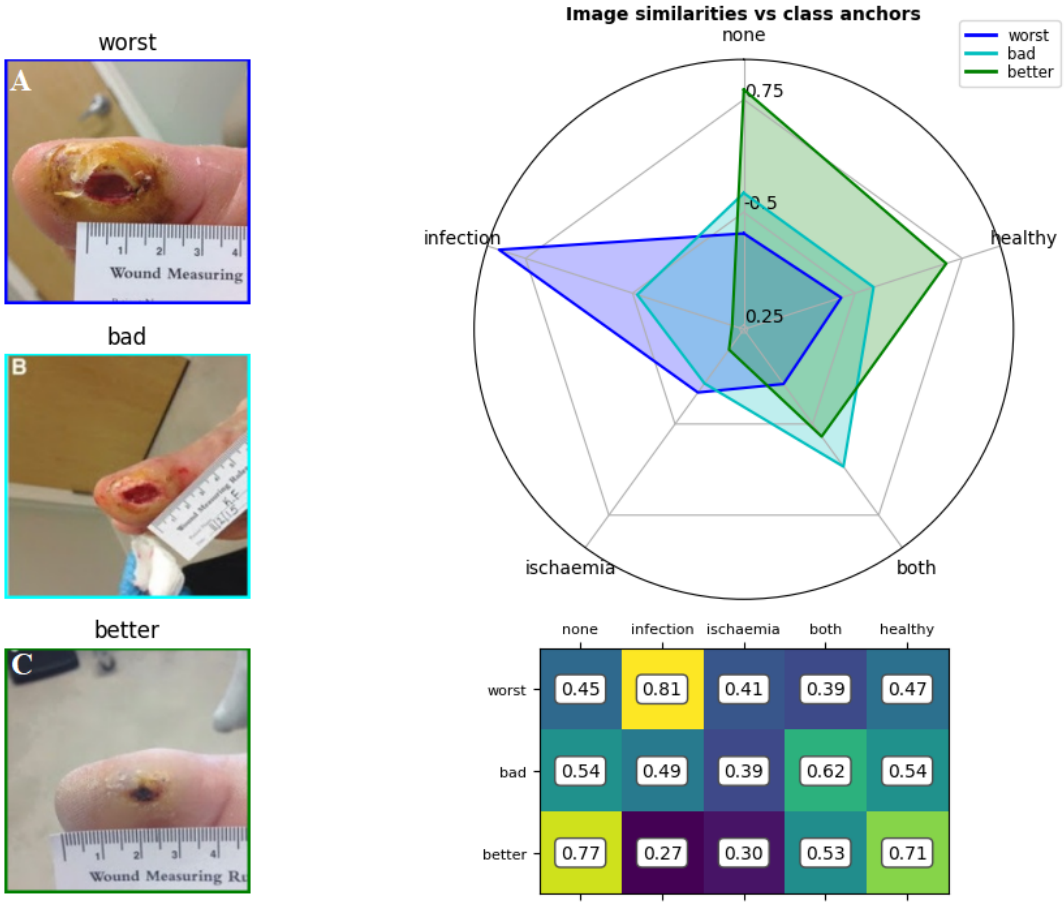


Figure 3.23: Longitudinal similarity for Use-Case #1: cosine similarity.

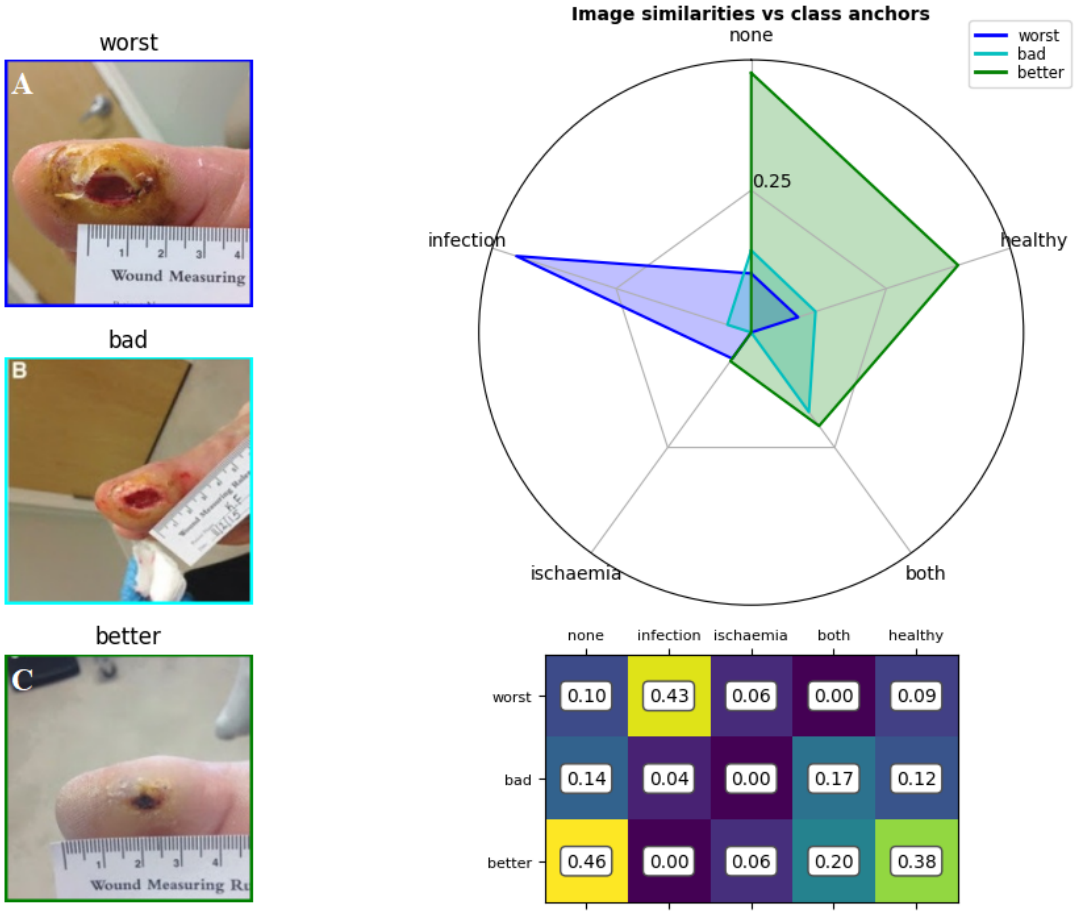


Figure 3.24: Longitudinal similarity for Use-Case #1: Euclidean distance.

Table 3.9: Results by applying cosine distance for Use-Case #1.

	None	Infection	Ischaemia	Both	Healthy
Image A	0.45	**0.81**	0.41	0.39	0.47
Image B	0.54	0.49	0.36	**0.62**	0.54
Image C	**0.77**	0.27	0.30	0.53	0.71

Table 3.10: Results by applying Euclidean distance for Use-Case #1.

	None	Infection	Ischaemia	Both	Healthy
Image A	0.10	**0.43**	0.06	0.00	0.09
Image B	0.14	0.04	0.00	**0.17**	0.12
Image C	**0.46**	0.00	0.06	0.20	0.36

Use-Case #2

For the second used case, we consider images from the work of Almonaci *et al.* [222], which is a chronological evolution of an ulcer in a diabetic male patient who is 54 years old with type 2 diabetes, with image A showing the initial appearance of the ulcer, image B showing the evolution of the DFU after 11 days, and image C after 12 days.

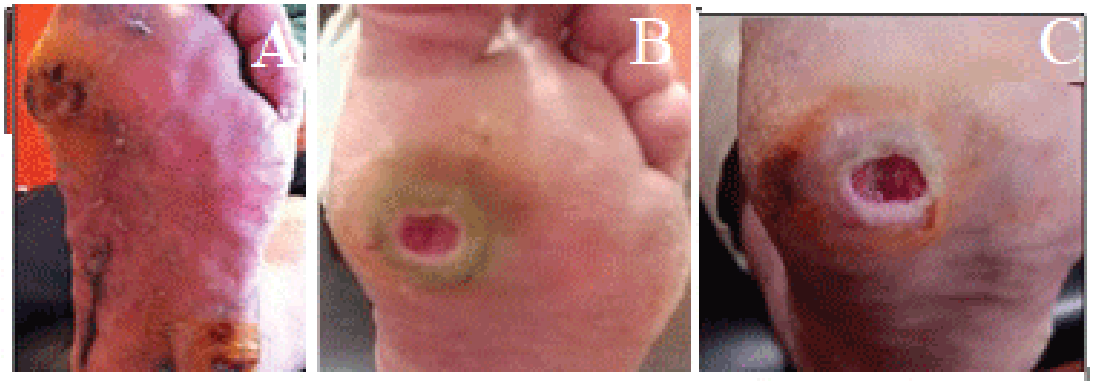


Figure 3.25: DFU images for Use-Case #2 [222].

From Figures 3.26 and 3.27, we can see that semantic similarity is closest to Infection on day 1 for image A. According to the process applied by the doctor after debridement of the lesion on day 11 [222], and the treatment protocol consisting of application of topical administration of AgNPs solution, we have the image after 12 days, which shows a drastic improvement as observed on both the radar charts, implying the ulcer is responsive to the treatment. However, based on the conclusion we studied in the work of Almonaci *et al.* [222], where they seem to be more satisfied with the result on day 13, our model tends to show a regression from day 12 to day 13 in both the radar plots. They tend to again point towards a higher similarity to infection. We shall now analyze what the numbers show in terms of similarity from Table 3.11 and Table 3.12. When the cosine distance is used on arrival on day 1, the similarity is clearly set to infection. However, once treatment is started, we can see that on day 12 there is a higher similarity to none and healthy, which is the same for Euclidean distance. For day 13, there is also the highest similarity to none, but instead of an increased similarity to both none and

healthy, it shows a slight reduction. There may be an influence on the way the images are taken, which opens up the need to have a clearly defined protocol for the DFU images.

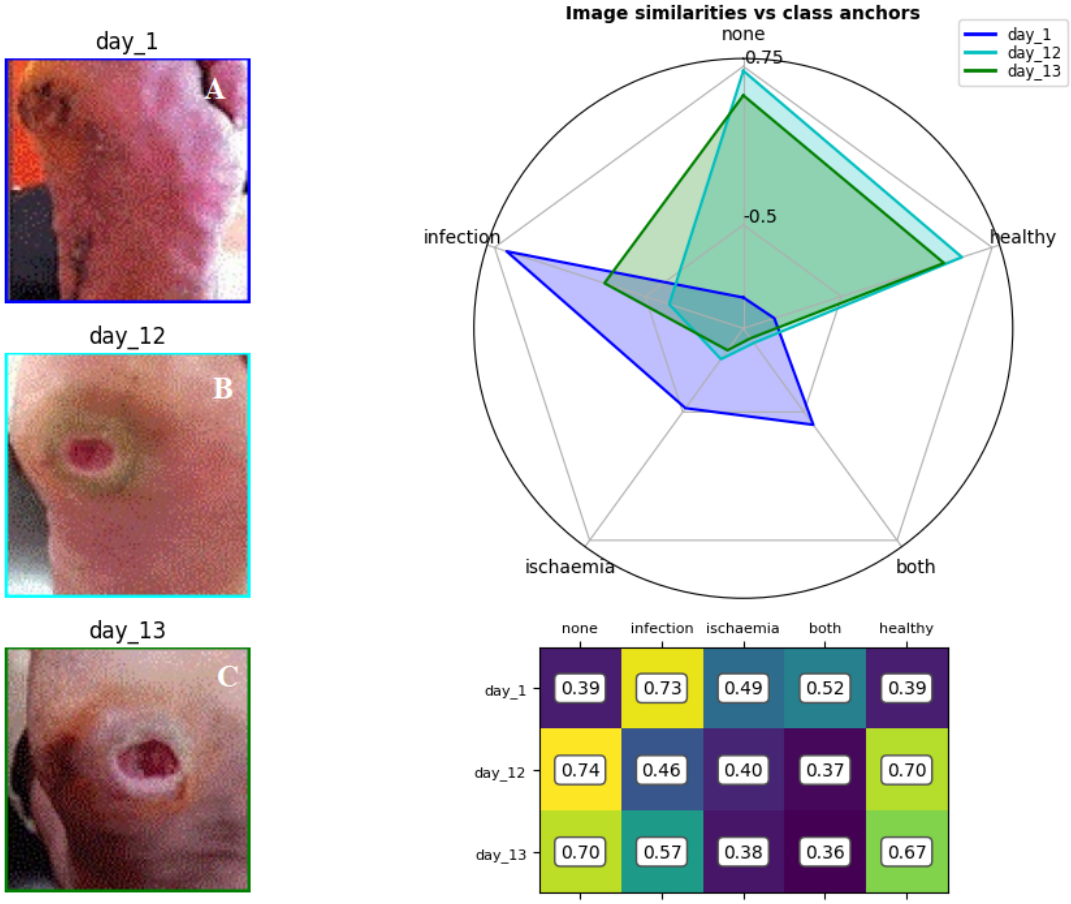


Figure 3.26: Longitudinal similarity for Use-Case #2: cosine similarity.

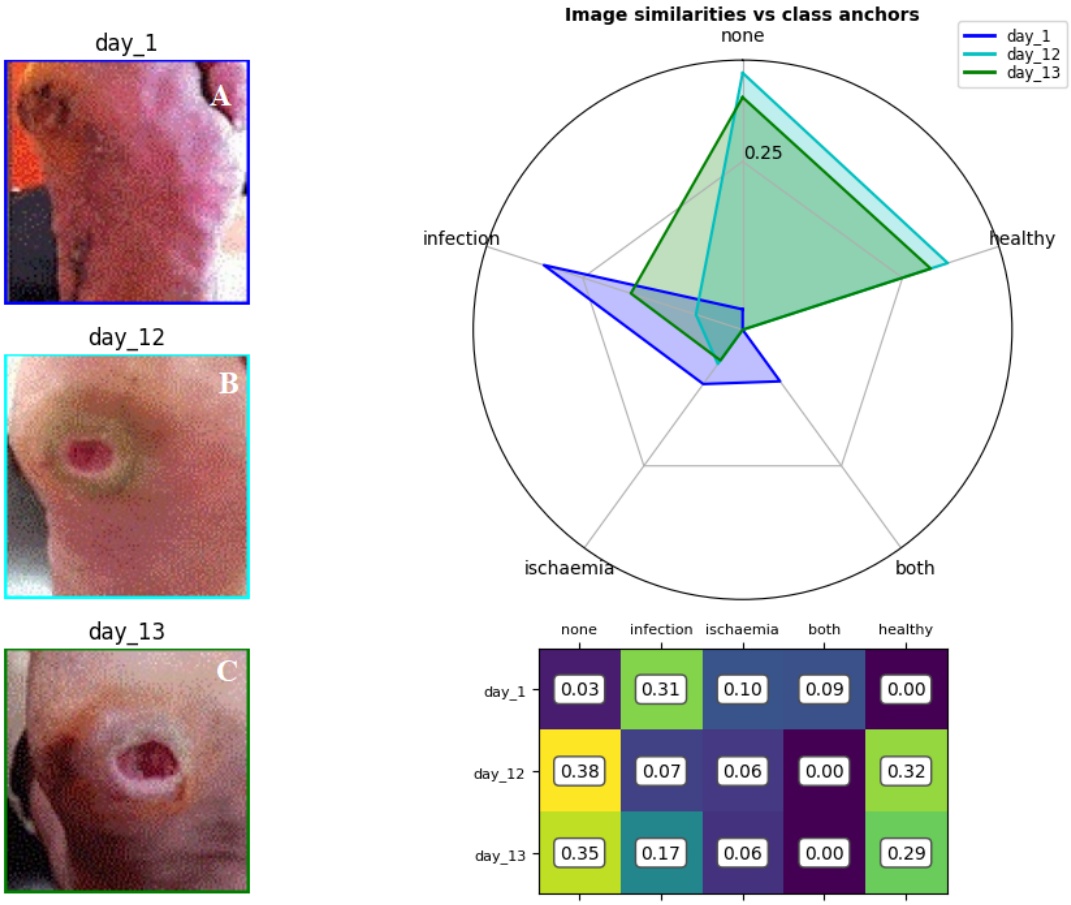


Figure 3.27: Longitudinal similarity for Use-Case #2: Euclidean distance.

Table 3.11: Results by applying cosine distance for Use-Case #2.

Time	Image	None	Infection	Ischaemia	Both	Healthy
Day 1	**Image A**	0.39	**0.73**	0.49	0.52	0.39
Day 12	**Image B**	**0.74**	0.46	0.40	0.37	0.70
Day 13	**Image C**	**0.70**	0.57	0.38	0.36	0.67

Table 3.12: Results by applying Euclidean Distance for Use-Case #2.

Time	Image	None	Infection	Ischaemia	Both	Healthy
Day 1	**Image A**	0.03	**0.31**	0.10	0.09	0.00
Day 12	**Image B**	**0.38**	0.07	0.06	0.0	0.32
Day 13	**Image C**	**0.35**	0.17	0.06	0.0	0.29

Use-Case #3

The images used in this third case were obtained from the study conducted by Tobalem *et al.* [223]. Figure 3.28 shows the progression of DFU over 10 days. The patient had photographed the lesion twice daily, thinking it would heal spontaneously (Panel A). The preoperative photographs show erythema (day 1), blisters (day 3), a necrotizing abscess (day 6), and a wound infection requiring surgery (day 10) [223]. We will use these images to show how the model and process we propose can be used not only by doctors but also by patients.

Figure 3.28: DFU images for Use-Case #3 [223].

The radar charts shown in Figures 3.29 and 3.30 offer valuable insights into the progression of DFU in the patient. On day 1, the model indicates a similarity of 0.6 with the infection class while also showing high values for the healthy and none classes, which may lead to some confusion. However, by day 2, the model clearly emphasizes the infection class, and subsequent days exhibit a shift towards both infection and ischaemia. These results effectively demonstrate the model's capability to track the development of DFU in the patient over time.

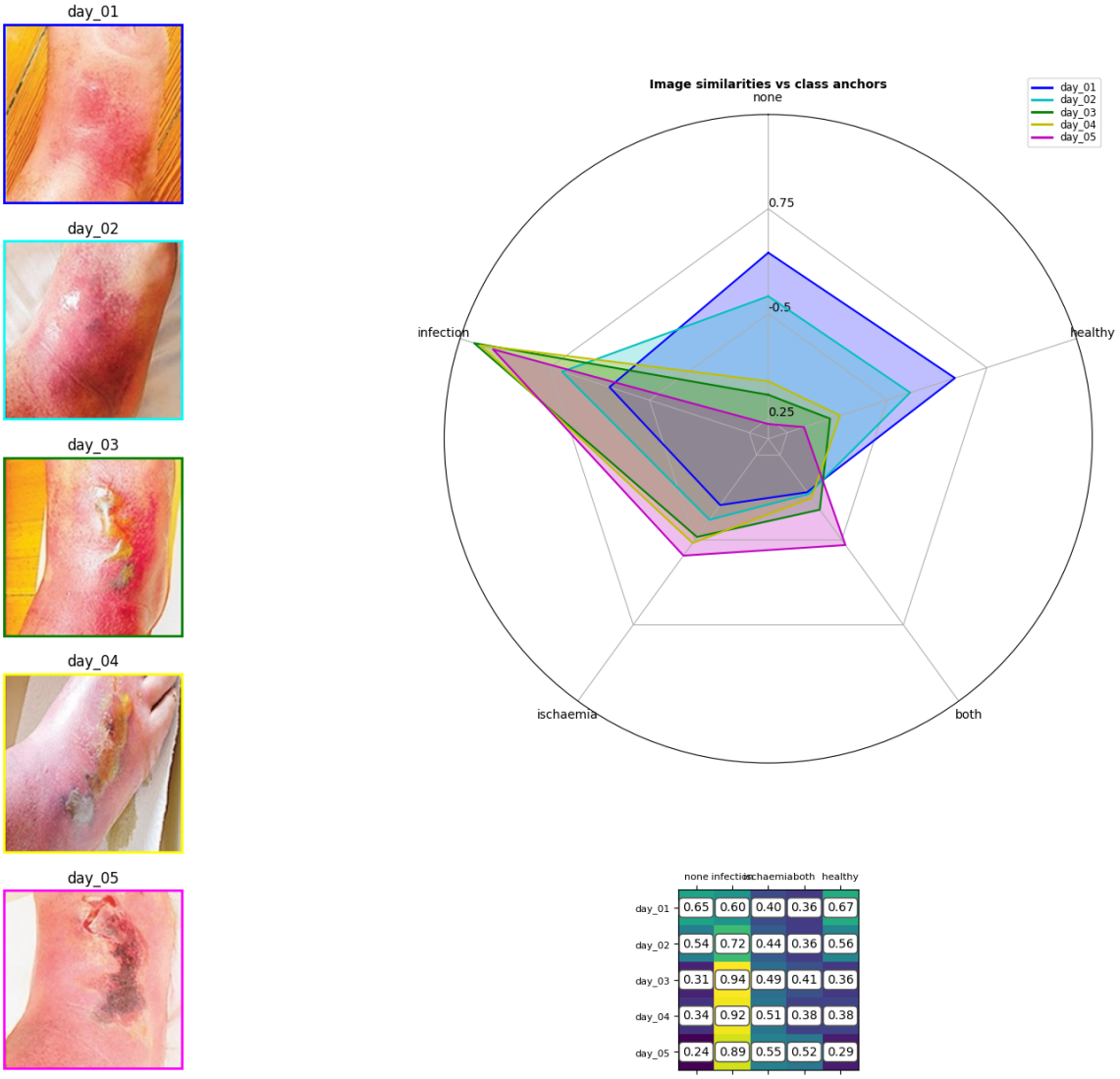


Figure 3.29: Longitudinal similarity: cosine similarity, part 1.

In this specific case, the model acts as supplementary evidence to illustrate the progression of DFU. Comparable conclusions can be drawn by examining Figures 3.31 and 3.32.
From Tables 3.13 and 3.14, the trends in terms of similarity follow the same pattern. From the research article [223], the following descriptions were obtained: erythema (day 1), blisters (day

3), a necrotizing abscess (day 6), and a wound infection requiring surgery (day 10). On day 10, with cosine distance, there is a high similarity between ischaemia and both, but for Euclidean distance, there is the highest similarity between both on day 10. However, from day 9 to day 10, the reduction in both similarities poses a problem, as based on the description from the research article, the situation is actually even more serious. Here, again, we see the need to have a control protocol for taking the picture or have our model better learn similarity by experimenting more in terms of architecture or dealing with the class imbalance.

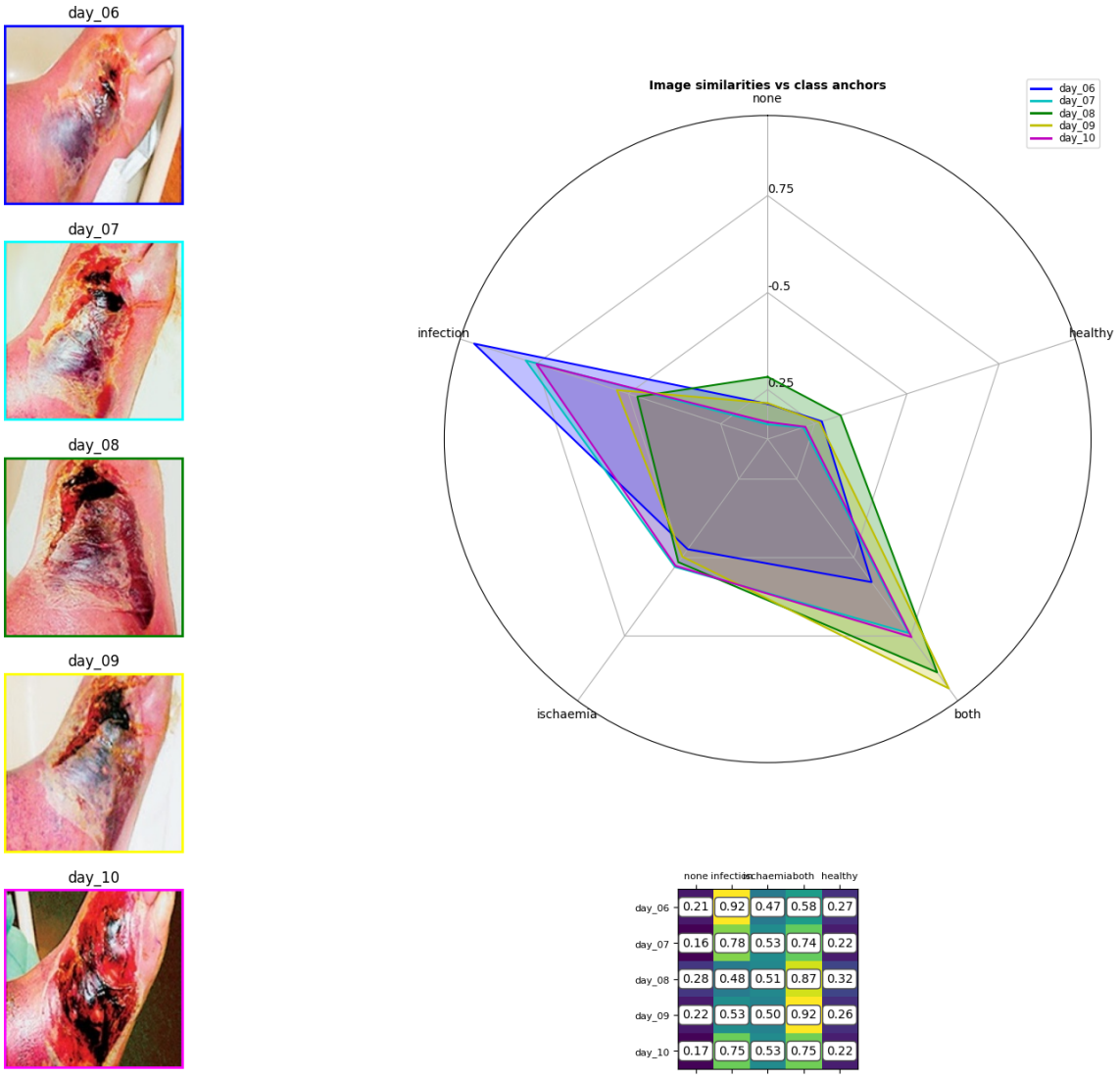


Figure 3.30: Longitudinal similarity: cosine similarity, part 2.

Table 3.13: Results by applying cosine distance for Use-Case #3.

Time	None	Infection	Ischaemia	Both	Healthy
Day 1	0.65	0.60	0.40	0.60	**0.67**
Day 2	0.54	**0.72**	**0.72**	0.36	0.56
Day 3	0.31	**0.94**	**0.94**	0.41	0.36
Day 4	0.34	**0.92**	**0.92**	0.38	0.38
Day 5	0.24	**0.89**	**0.89**	0.52	0.26
Day 6	0.21	**0.92**	0.47	0.58	0.27
Day 7	0.16	**0.78**	0.53	**0.74**	0.22
Day 8	0.28	0.48	0.51	**0.87**	0.32
Day 9	0.22	0.53	0.50	**0.92**	0.26
Day 10	0.17	**0.75**	0.53	**0.75**	0.22

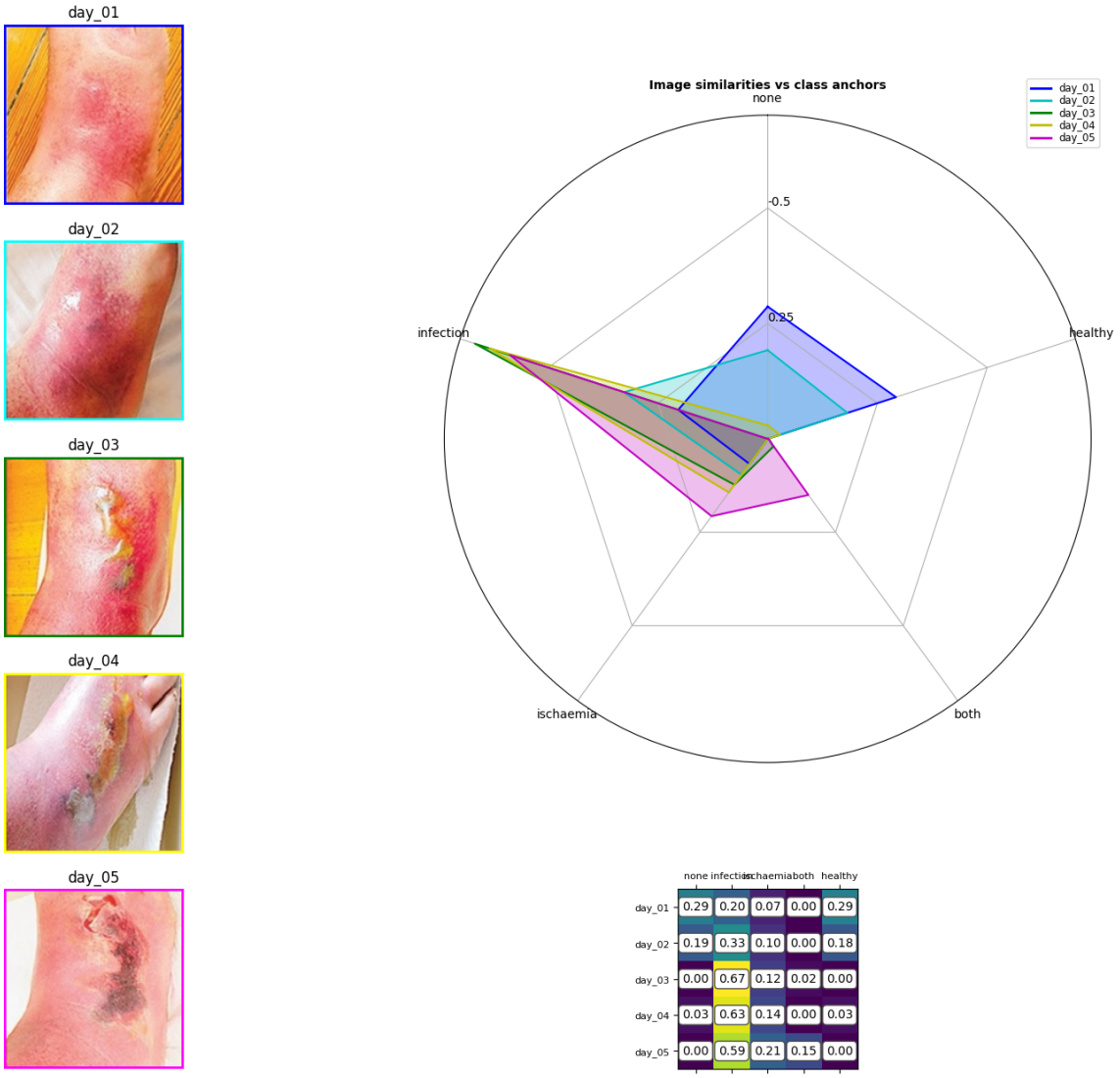


Figure 3.31: Longitudinal similarity: Euclidean distance, part 1.

Table 3.14: Results by applying Euclidean distance for Use-Case #3.

Time	None	Infection	Ischaemia	both	Healthy
Day 1	**0.29**	0.20	0.07	0.00	**0.29**
Day 2	0.19	**0.33**	0.1	0.00	0.18
Day 3	0.00	**0.67**	0.12	0.02	0.00
Day 4	0.03	**0.63**	0.14	0.00	0.03
Day 5	0.00	**0.59**	0.21	0.15	0.00
Day 6	0.00	**0.64**	0.16	0.22	0.00
Day 7	0.00	**0.44**	0.23	0.40	0.00
Day 8	0.01	0.09	0.16	**0.53**	0.00
Day 9	0.00	0.18	0.18	**0.64**	0.00
Day 10	0.00	0.22	0.22	**0.41**	0.00

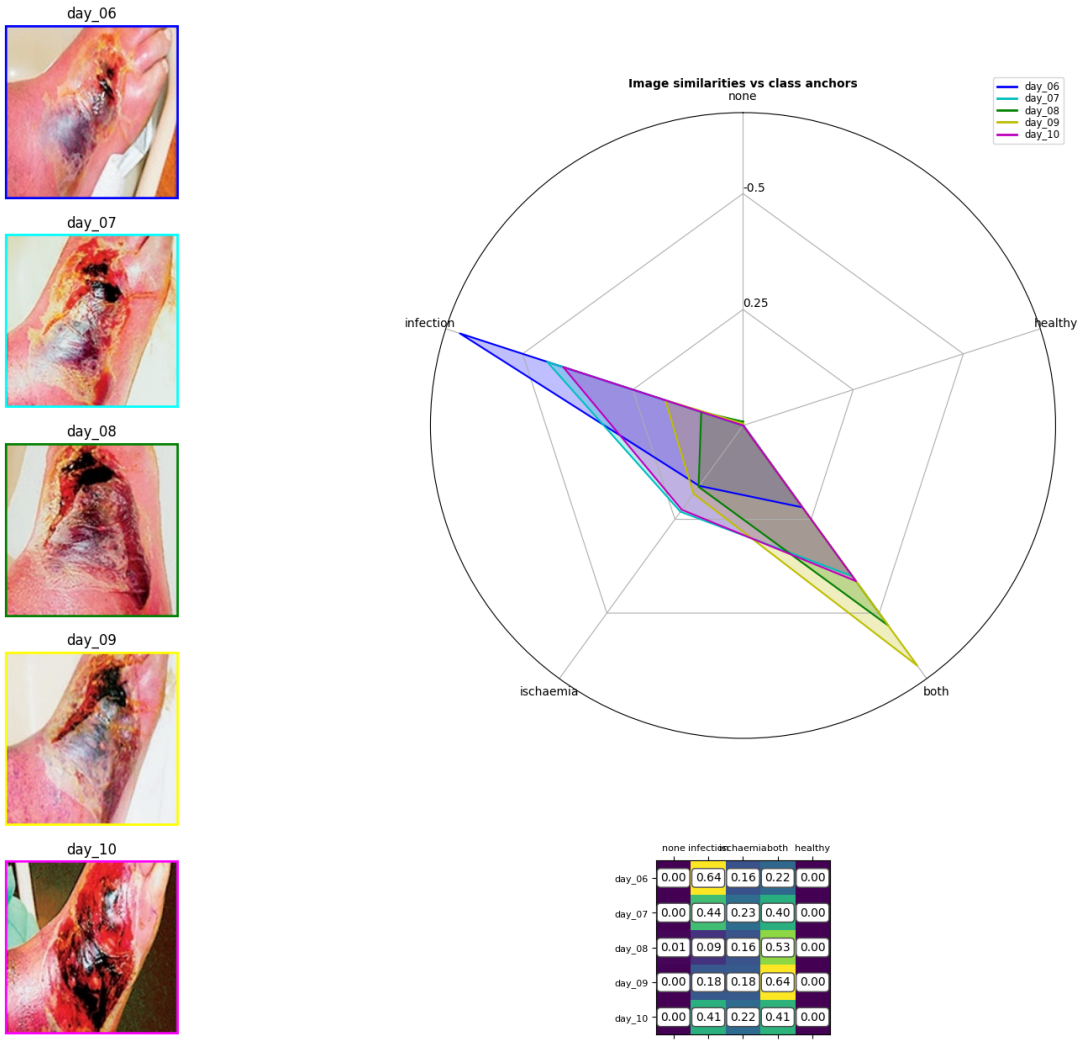


Figure 3.32: Longitudinal Similarity: Euclidean distance, part 2.

3.3.5 Discussions

DFU-Helper is a framework that utilizes Siamese Similarity Learning to assist medical practitioners in validating follow-up diagnostics for treatment purposes. While our model demonstrates a strong correlation with the evaluation conducted in the research articles from which we sourced the test images, we were unable to assess its performance on real longitudinal DFU images, as was done by Li *et al.* [201] in their studies on retinopathy of prematurity and osteoarthritis. Li *et al.* [201] had a database of longitudinal images consisting of 4861 images from 870 patients for retinopathy of prematurity, and 10,012 images from 3021 patients for osteoarthritis. In contrast to our approach of showing progress across five class anchors, Li *et al.* [201] employed an overall severity score and compared individual images against the established normal standards for both eyes and knees, providing a corresponding score as output.

As far as the architecture of our SNN is concerned, as opposed to all the work detailed in the related works sections, including those for severity evaluation over a time period [201], [203], [204], none uses an ensemble of CNN and ViT for the sub-networks. For CNN, we use EfficientNet, which is known to achieve state-of-the-art results while dealing with images, the novel and widely used ViT, and more specifically, the BEiTV2 [178]. When tested on a classification task, this architecture has a Macro F1-score of 0.951 and an accuracy of 0.9395. The majority of work are concentrated on using CNN backbones namely VVG16, ResNet18, and InceptionV3. In the study conducted by Ornob *et al.* [208], they developed a detection method for COVID-19 by creating an ensemble of six pre-trained CNNs and the Vision Image Transformer (ViT), specifically the Swin Transformer, along with a Triplet Siamese Neural Network framework. Although we have concerns about the computational resources required, the concept of integrating cutting-edge models to achieve optimal results in the medical field is acceptable. As part of our future work, we propose evaluating the top-performing ensemble CNN and the top-performing ensemble ViT separately and subsequently combining the best ensembles to further improve detection accuracy.

In the selection of anchors for image comparison over time in DFU-Helper, we took a different approach compared to AbdulRaheem *et al.* [203] and Akbar *et al.* [204]. Instead of using 5 or 16 images per class, we utilized the maximum number of available images per class in our dataset to generate the class anchors. For training DFU-Helper and SNN, we introduced a novel loss function called Large Margin Cotangent (LMCoT), proposed by Duong *et al.* [173], deviating from the commonly used contrastive loss or triplet loss functions in the relevant works we considered for severity estimation, disease detection, or classification. Interestingly, Akbar *et al.* [204] explored four different loss functions, including contrastive loss, mean square error (MSE) loss, Huber loss, and a combination of contrastive and MSE loss, which holds its own merit.

In contrast to non-deep learning methods, Hu *et al.* [206] simplifies the computational complexity by converting images to grayscale during processing. However, in our DFU-Helper, the SNN analyzes images while retaining the three RGB channels. Moreover, we used meticulously labeled data to ensure accurate training and validation for similarity learning in our model. The mentioned work does not specify the number of experts whose input was considered to calculate similarity, which contrasts with our algorithm that outputs similarity results. Additionally, in comparison to the approach proposed by Ionescu *et al.* [207], which solely focuses on image equivalence without considering image content, our SNN, particularly the utilization of the ViT, places significant emphasis on the semantic content of the image.

The DFU-Helper framework can further be enhanced by incorporating additional deep learning architectures. To ensure its suitability for application in the public health domain, it is essential to establish a rigorous validation and testing protocol, thereby making the proposed framework more robust.

3.3.6 Limitations

One key limitation of our study is the quality of the dataset. With the promising results that deep learning is showing for exploiting medical images, there is a need to have a better-quality dataset. In this case, we had to introduce the healthy class images from another source. Secondly, with the limitation in processing power, there are other ensemble models for the sub-network that we were not able to test and evaluate, which could have given a better result in terms of showing similarity or dissimilarity. One of the disadvantages of the Siamese Neural Network is the high processing power needed. Finally, we do not have a way to compare the output of the DFU-Helper framework with a known severity grading, as in the case of the work of Li *et al.* [201].

3.3.7 Conclusions and Future Works

In this research, we used similarity learning to train an SNN and propose a DFU-Helper framework to assist doctors who diagnose and decide on the treatment protocol of DFU, and subsequently perform follow-up on a patient. We combined the CNN and ViT transformers and used transfer learning on the DFU dataset. As exposed in the limitations, there are several other models and techniques that can be further experimented with.

Machine learning offers a set of techniques and methods that can turn raw data into realistic and tailored knowledge [224] which can, in turn, give additional insight to healthcare professionals. This tool must undergo testing in collaboration with healthcare experts who can formulate suitable protocols for its implementation within a cohort, as well as the collection of performance indicators.

In our study, our intention was not to replace healthcare professionals but rather to offer an additional perspective through the utilization of deep learning and image similarity architecture. However, it is important to highlight that the tool represents a novelty in the realm of DFU management. Currently, the management heavily relies on the practitioner's experience, which, regrettably, can be susceptible to misinterpretation due to their workload. Incorporating DFU-Helper provides confirmation of their diagnoses, ultimately enhancing the reliability of the tool. Upon completing this phase, the ultimate objective would be to package the same functionality into a mobile application, enabling patients to engage in self-monitoring.

The prompt detection and efficient management of diabetic foot ulcers can play a crucial role in preventing the advancement of wounds and reducing the risk of amputations. Our research has the potential for further development as a tool for early detection of DFUs, as demonstrated by our experimental findings with the addition of the new Healthy Class.

An essential aspect of future work in this field is involving healthcare institutions in the data collection process, as data remains crucial for model development. This will necessitate

further research into data confidentiality within the context of machine learning, and exploring federated learning models could be a potential avenue.
Secondly, the incorporation of explainability into the system is imperative. Subsequent research needs to focus on ensuring that the output of DFU-Helper is interpretable, thus gaining acceptance from medical professionals. This avenue opens up the realm of Explainable Artificial Intelligence (XAI), an area that has gained significant momentum in recent times.
The medium-term goal remains the initial deployment for doctors, with the ultimate aim of creating a patient application. This approach is aligned with the research conducted by Plonderer *et al.* [225], which emphasizes the utilization of such tools in collaboration with healthcare experts.

4

Confidentiality of Healthcare Data

Summary

4.1 Introduction

Recent advancements in the field of machine learning (ML) have led to highly effective innovations across various domains. For instance, in the specialized field of dermatology, a machine learning model has been employed to diagnose skin cancer and has achieved comparable results to dermatologists [226], [227]. This is true when dealing with digital medical images as well as dealing textual health data with the possibility of generating reports which extracts quantitative, objective, structured, and personalized information from stroke MRIs with performance comparable to that of an expert evaluator.
Furthermore, many recent ML applications rely heavily on deep learning [228], which necessitates sufficiently large and diverse datasets to ensure reliability [229]. However, the collection of such datasets can be challenging. In many domains, data are owned by numerous clients and are stored in various locations. Due to privacy and regulatory concerns, data sharing among clients is not possible. The issues associated with data sharing make it difficult to generate robust ML models. Consequently, the existing collected data is not fully leveraged by ML. This unfortunately negatively impact the development of high performing ML models. Robust ML models have the potential to enhance efficiency and reduce costs in numerous fields and one such field which is of concern to us in this study is healthcare [230], [231].
Federated learning is a method that allows clients to collaboratively train ML models without sharing raw training data [232]. Normally, ML models are trained in a centralized location where the model owner can freely observe all the training data. However, in federated learning (FL), model training is decentralized. The predominant FL strategy involves using a central orchestration server that distributes a global model to participating clients. These clients then train the models using their local data. The updated parameters of the local model are then sent to the central server, where the global model is updated by aggregating and combining the parameters from the clients' models. In the industry, some large technology companies have adopted FL in production, and many startups intend to use FL to address regulatory and privacy concerns. However, FL poses several challenges, such as communication efficiency, system heterogeneity, non-identically distributed client data (non-IID), and privacy protection [233]. For example, non-IID client data, such as an imbalanced distribution of labels, can significantly impede the learning process [234].

With centralized FL, clients must trust and rely on a central server. This approach carries the risk of disrupting the training process in case of server failure. Additionally, in FL scenarios where the number of participating clients is potentially high, the central server must handle a large number of communications, which can be a limiting factor [234]. To address certain issues associated with centralized FL, peer-to-peer (P2P) FL could be a viable alternative as it allows for bypassing dependence on the central server. To achieve this, we extend an important centralized FL algorithm called FedAvg [232] to operate in a P2P environment. This extension draws inspiration from other works that explore decentralized model training [235], [236], [237]. We will investigate P2P FL by training deep neural networks using the dataset obtained from the DFU 2021 [195] challenge organized by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) society [170] which is used for the classification of Diabetic foot ulcer (DFU).
It should be noted that our work draws inspiration from Mäenpää's study [238]. However, we have expanded upon this foundation by employing it in the context of a tangible medical dataset and on a novel deep learning architecture. Additionally, we extended our investigation by incorporating P2P federated learning using the Stochastic Gradient Descent (SGD) approach. The main objective of this work is to study the feasibility of FedP2P on our model and dataset. To achieve this, we start by extending the FedAVG algorithm [232] to operate in a P2P environment. Two types of algorithms, FedAVGP2P and FedSGDP2P is studied. A comparative evaluation is conducted between FedAVGP2P and FedaAVG, considering empirical assessments. The aspects taken into consideration include model convergence behavior and communication costs. These aspects will be examined in scenarios involving both IID and non-IID local client data. Furthermore, this work aims to explore possible means of improving FedAVGP2P and FedSGDP2P by implementing specific heuristics.
The subsequent sections of this chapter are organized as follows: Section 4.3 provides a comprehensive analysis of relevant previous works in the field. In Section 4.4, we present a detailed description of our proposed solution. This is followed by the presentation of our experimental results in Section 4.5. Finally, we conclude the paper with a summary of our findings and suggestions for future research, which are presented in Section 4.6.

4.2 Background and Preliminaries

In this section we introduce and explain the reason for federated learning. We also describe the two distinct Federated Learning architectures: Centralised Federated Learning and Peer-to-Peer Federated Learning architectures.

4.2.1 Federated Learning (FL)

FL, also referred to as collaborative learning, represents a decentralized method for training machine learning models. Federated learning is a distributed machine learning technique that trains a global model by exchanging model parameters or intermediate results among multiple data sources [239]. In this technique, data are not transferred from client devices to centralized servers. Instead, the raw data residing on ubiquitous devices is utilized for local model training, thereby enhancing data privacy. The ultimate model is collaboratively constructed by combining these local updates.
FL is an attractive avenue to explore when dealing with sensitive data which are not located at a centralised server as it ensures:

1. Privacy and confidentiality: federated learning allows for training to occur locally on the edge device, preventing potential data breaches during transfer.
2. Data Security: Only the encrypted model updates are shared with the central server, assuring data security.
3. Access to large scale data which can reduce data bias. Federated learning guarantees access to data spread across multiple devices, locations, and organizations. This can ensure data diversity, which in turn ensures that models can be made more generalizable.

4.2.2 Centralised Federated Learning Architecture

In Centralised Federated Learning there exists a centralised server which coordinates the whole training process.
The central server is responsible of the following task:

1. Determines a global model to be trained
2. Selects participants (i.e., local nodes) for each training round
3. Aggregates local training results sent by the participants
4. Disseminates the updated model to the participants
5. Terminates the training when the global model satisfies some requirements (e.g., accurate threshold is reached).

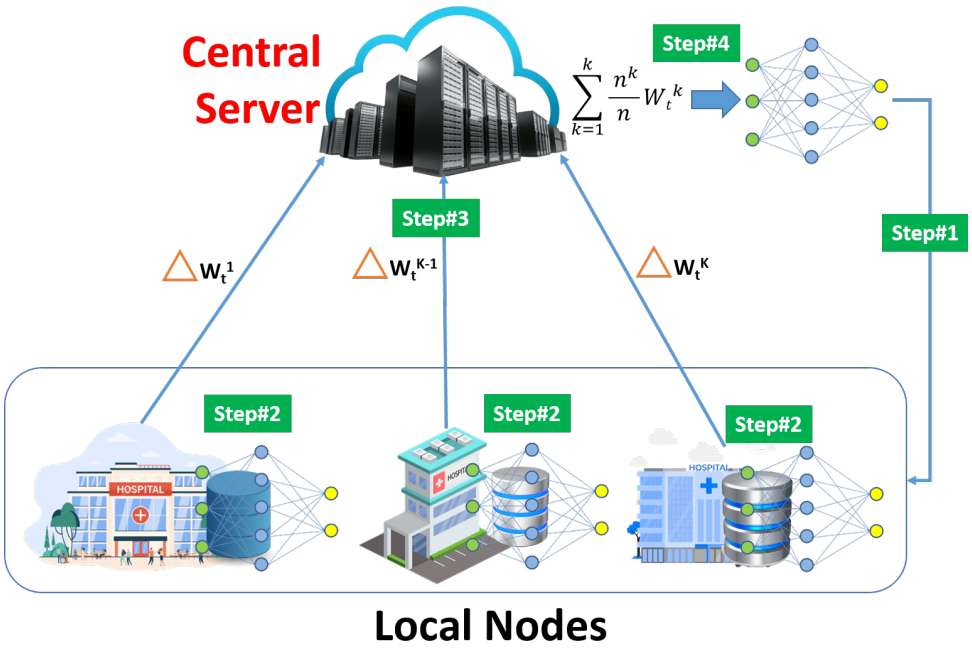


Figure 4.1: The centralised FL Architecture Inspired by [240] Step1: Participant Selection and Global Model Dissemination. Step2: Local Computation. Step3: Local Models Aggregation. Step4: Global Model Update.

Algorithm 4.2 shows the mechanics of the centralised architecture. From the network perspective we can immediately deduce that this architecture generates high-communication cost between server and clients and is also a vulnerable point of failure of the overall learning process.

4.2.3 Federated Learning: Peer-to-Peer Architecture

The architecture of Federated Learning based on peer-to-peer interaction operates without the need for a central server to coordinate the learning and parameter sharing process. Participants engage in direct communication without relying on an intermediary and this results in an equitable standing for each participant within the architecture, enabling any participant to initiate a model exchange request with others [241]. Due to the absence of a central server, participants must establish a prior consensus regarding the sequence in which models are to be transmitted and received.

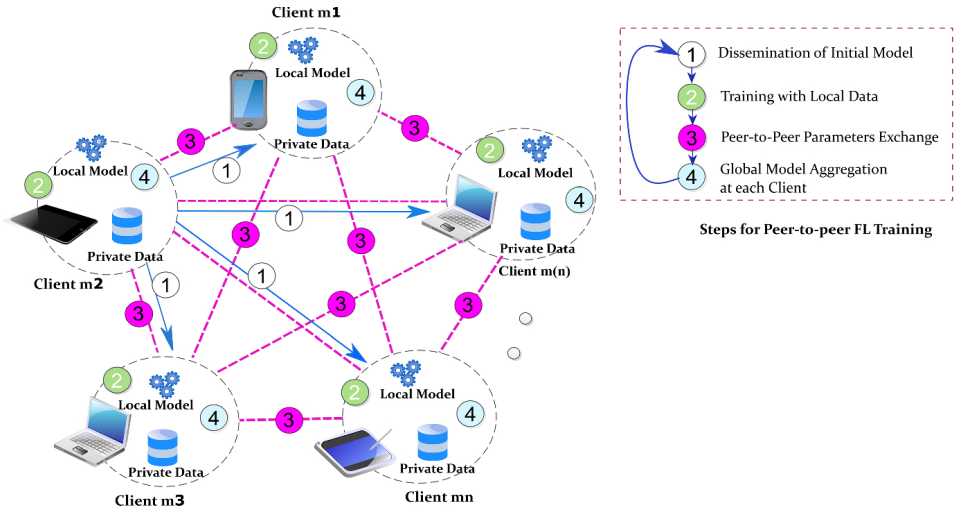


Figure 4.2: The P2P FL Architecture inspired by [242] clients directly communicate with one another instead of any central authority. A group of clients with a common goal collaborate to improve their models by sharing information from peer to peer.

When assessing vulnerabilities, the P2P FL architecture proves superior due to its avoidance of a central server, mitigating the risks associated with a single point of failure. Nonetheless, the efficiency of this approach can be influenced by the manner in which clients are interconnected [243], potentially impacting communication costs. Hence, achieving an equilibrium between performance and communication expenses becomes imperative within the P2P FL framework.

4.2.4 Federated Learning Algorithm

In federated learning, an aggregation algorithm refers to a technique implemented for consolidating the outcomes of training numerous intelligent models on the clients' devices, utilizing their respective local datasets. This algorithm plays a crucial role in combining the results derived from the local client training processes and subsequently updating the global model [244]. Two such algorithms are:

1. Federated stochastic gradient descent (FedSGD) averages the locally computed gradient at every step of the learning phase.
2. Federated averaging(FedAVG) averages local model updates when all the clients have completed training of their models.

Before moving forward, we shall introduce some terms:

1. Round: A round in federated learning is an iteration of the federated learning process. In each round, a subset of clients is selected to participate in the training process.
2. Clients: k is randomly selected subset of K clients to participate in the current epoch.
3. Non-IID dataset: This stands for non-independent and identically distributed dataset. For an image classification problem is means we may have some classes which exists at some clients while they do not exist at other clients. Non-IID poses a challenge to deep

learning models as it can lead to biased or unreliable models, resulting in low accuracy and incorrect results.

4. IID Datatset: This stands for independent and identically distributed dataset. For an image classification problem, it means that each image has a similar probability distribution as the others, and all are mutually independent.

Federated stochastic gradient descent (FedSGD)

FedSGD is an optimization algorithm used in Federated Learning (FL) to train machine learning models on decentralized data. It is a variation of the traditional Stochastic Gradient Descent (SGD) algorithm, adapted to the federated setting. FedSGD is a distributed version of SGD and uses the computation power of several compute nodes instead of one [245]. In FedSGD [246], the central model is distributed to the clients, and each client computes the gradients using local data. These gradients are then passed to the central server, which aggregates the gradients in proportion to the number of samples present on each client to calculate the gradient descent step. The key difference between FedSGD and traditional SGD lies in the aggregation step. In SGD, the local updates from all devices are typically averaged to update the global model. Moreover, a fraction of devices is randomly selected to participate in each round of model updates. This selective participation helps to reduce communication overhead and computational burden.

Algorithm 4.1 Federated Stochastic Gradient Descent (FedSGD) Algorithm

Input:
Global model parameters: θ_0
Number of federated rounds: T
Learning rate for clients: η
Initialization:
Initialize global model parameters: θ_0
for $t = 1$ to T **do**
    Select a subset of client devices: $\mathcal{C}_t$
    for each client $i \in \mathcal{C}_t$ in parallel **do**
        Receive the current global model parameters: θ_{t-1}
        Sample a mini-batch of local data: $\mathcal{B}_i$
        Compute the local gradient: $g_i \leftarrow \nabla f_i(\theta_{t-1}; \mathcal{B}_i)$
        Update the client's local model: $\theta_i^t \leftarrow \theta_{t-1} - \eta \cdot g_i$
    Aggregate local models to update the global model:
    $\theta_t \leftarrow \sum_{i \in \mathcal{C}_t} \frac{|\mathcal{B}_i|}{|\mathcal{B}|} \cdot \theta_i^t$
Output: Final global model: θ_T

Since there is the need to send parameters to main server after each gradient calculation has a bandwidth cost with may be a problem if the clients have limited connectivity access. This issue is tackled by Federated averaging (FedAVG).

Federated averaging (FedAVG)

Federated averaging (FedAVG) is a communication-efficient algorithm for distributed training with an enormous number of clients [247]. It ensures data privacy and security and maintains data locality by enabling model training without sharing the raw data. It uses one aggregation by the server in each communication round, which significantly reduces the communication cost between server and clients. Instead of sharing the gradients with the central server, weights tuned on the local model are shared. Finally, the server aggregates the clients' weights (model parameters). The fundamental idea is that clients run multiple updates of model parameters before passing the updated weights to the central server [245].

Algorithm 4.2 FederatedAveraging. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate [232]. C is the fraction of participating devices.

Server executes:
Initialize w_0
for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C, K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ in parallel **do**
        $w_{t+1}^k \leftarrow$ (ClientUpdate) (k, w_t)
    $w_{t+1} \leftarrow \sum_{k+1}^{K} \frac{n_k}{n_t} w_{t+1}^k$
function ClientUpdate$((k, w))$ ▷ *Run on client k*
    $\beta \leftarrow$ (split P_k into batches of size B)
    for each local epoch $i = 1$ to E **do**
        for each batch $b \in \beta$ **do**
            $w \leftarrow w - \eta \nabla l(w; b)$
    return w to server

Federated averaging Peer-to-Peer(FedAVGP2P)

FedAVGP2P is an extension or variation of the FedAVG algorithm in federated learning. In FedAvg, a central server coordinates the model aggregation process, where local models from participating clients are averaged to update a global model.
In the FedAVGP2P variant, the aggregation process involves peer-to-peer communication among participating clients, bypassing the need for a central server. Clients directly communicate with each other to exchange their local model updates, and collectively compute the global model through decentralized means.

Federated SGD Peer-to-Peer(FedSGDP2P)

FedSGDP2P is an extension or variation of the Federated Stochastic Gradient Descent (FedSGD) algorithm in federated learning. In the standard FedSGD algorithm, a central server coordinates the federated learning process, where clients compute gradients on their local data and send them to the server for aggregation and model updates. In the FedSGDP2P variant, the communication process occurs directly between participating clients in a peer-to-peer manner, eliminating the need for a central server. Clients collaborate with each other to exchange gradient information and update their models collectively. This approach has the potential to enhance privacy, reduce communication overhead, and improve the scalability of federated learning. However, it may introduce challenges related to synchronization, security, and the management of peer-to-peer networks.

4.3 Related Work

This section investigates the primarily application federated learning for the confidentiality of data.

In [244], moshawrab *et al.* reviews the use of federated learning and its application in the prediction of disease. They discuss the use of FL for diagnosing FL in the diagnosis of cardiovascular disease, diabetes, and cancer. Quite naturally with the use of medical data they stress on the need for privacy and confidentiality in dealing with sensible data. They identify other areas, beside healthcare, where implementation of FL makes sense including Smart retail, Transportation, Natural language processing and Finance.

When dealing with FL there is the need to strike a balance between performance and communication cost. Asad *et al.* [248] consequently evaluated the cost of communication efficiency in a FL algorithm. They relied on latency and bandwidth as limitation and proposed the use of Averaging Algorithm (FedAVG) and Sparse Ternary Compression (STC), Communication-Mitigating Federated Learning (CMFL), Federated Maximum and Mean Discrepancy (FedMMD) to evaluate communications efficiency. All the algorithms were evaluated on CIFAR and MNIST dataset and using a model which is Convolution Neural Network (CNN) based. The data were divided in two ways to cater for Independent and Identically Distributed (IID) scenario and non-IID. The following parameters were used in the evaluation, client=100, number of classes= 10, batch size=20 and participating=10%. Unfortunately, in this research none of the algorithms were able to prove to be the best solution. However, the authors use this work to identify gaps and provide avenues for future research.

He *et al.* [235] introduced COLA, a decentralized training algorithm designed to optimize communication efficiency, scalability, and elasticity, while also accounting for unreliable and heterogeneous devices to accommodate data changes while Lin *et al.* [249] explored approaches

for enhancing mini-batch stochastic gradient (SGD) algorithms and presented a novel post-local SGD method that achieves remarkable performance gains compared to training with large batches. These improvements were observed across well-known benchmark datasets, all while ensuring efficiency and scalability. Roy *et al.* [250] introduced a fully decentralized architecture called P2P FL (Peer-to-Peer Federated Learning) to overcome the limitations of classical federated learning. The conventional federated learning approach involves a centralized controller that collects and consolidates training data from all nodes, maintaining a global model on a cloud-based infrastructure across the network. However, the P2P FL architecture deploys nodes throughout the network, allowing them to interact exclusively with their immediate neighbors, thus eliminating the necessity of a centralized controller. This development in P2P federated learning enables nodes to engage with their next hop neighbors in just two steps.

4.4 Proposed Methods

In FedAVG, a centralised server is mandatory for taking care of all transactions. By referring to from previous research on decentralized training algorithms [235], [236], [237], we enhance FedAVG to operate within a peer-to-peer framework, thereby eliminating the necessity for a central server. We furthermore extend our study by applying another variation of Federated Learning which is FedSGD [251].
The extended algorithms are referred to as FedAVGP2P and FedSGDP2P. Each client has their own model and communicate directly to other clients. Before training, all client models are initialized with the same weights. Each client performs training of the model using its local data. Then, each client aggregates and averages updates from a set of random neighbors or selected using a heuristic. This process is repeated for a finite number of rounds, allowing each client to have a fully trained global model without relying on a central server. A similar distributed computation is performed by the FedSGDP2P algorithm: during each round, clients calculate the gradient derived from the loss function on their local data. These gradients are then sent to other selected clients (either randomly or based on heuristics) to aggregate them and update the parameters of their models. Similar to FedAVG, FedAVGP2P and FedSGDP2P have four hyperparameters: the fraction of neighbors from which each client receives updates, the size of the local minibatch, the number of times each client trains on the shortest time period, the number of times each client trains on the local dataset in each round (epochs), and the learning rate.

4.4.1 Heuristic 0: random

This approach is done in a naive manner where we simply perform random sampling. In other words, each client will randomly send its weight/vector gradient to a subset of other clients.

FedAVG

Algorithm 4.3 FedAVG heuristic 0: random. c is fraction of clients that perform computation on each round

```
for round ← 1, 2, 3, ... do
    for client ← 1, 2, 3, ... do
        w_client = fit(w_client, data_client, epochs=$local_epoch)
    for client ← 1, 2, 3, ... in parallel do
        w_client ← Mean(GetRandomNeighbors(c).weight)
```

FedSGD

Algorithm 4.4 FedSGD heuristic 0: random. c is fraction of clients that perform computation on each round

```
for round ← 1, 2, 3, ... do
    for local_epoch ← 1, 2, 3, ... do
        for step ← 1, 2, 3, ... do
            for client ← 1, 2, 3, ... in parallel do
                grad_client = Gradient(w_client, data_client)
                grad = getRandomNeighborsGrad(c)
                w_client+ = Mean(grad)
```

In the original FedAVGP2P algorithm, the selection of neighbors for communication is done randomly. In order to enhance the performance of FedAVGP2P, we propose three distinct heuristics for choosing the neighbors to communicate with.

4.4.2 Heuristic 1: n lastest

Each client in the network maintains its own identity and keeps track of the identities of the n most recent clients it has interacted with. At the end of each communication round, this information regarding the n most recent clients is disseminated throughout the network. Subsequently, each client selects its communication partners based on the level of dissimilarity in their previous interactions. Specifically, clients prioritize communication with those who have had the least amount of overlap in past interactions.

FedAVG

Algorithm 4.5 FedAVG heuristic 1: n lastest. c is fraction of clients that perform computation on each round

for round $\leftarrow 1, 2, 3, \ldots$ **do**
  for $client \leftarrow 1, 2, 3, \ldots$ **do**
    w_{client} = fit(w_{client}, $data_{client}$, epochs=$local_epoch)
  for client $\leftarrow 1, 2, 3, \ldots$ in parallel **do**
    neighbors = GetRandomNeighbors(c, without = client.last)
    $w_{\text{client}} \leftarrow$ Mean(neighbors.weight)
    client.last = (client.last + neighbors)[-n:]

FedSGD

Algorithm 4.6 FedSGD heuristic 1: n lastest. c is fraction of clients that perform computation on each round

for round $\leftarrow 1, 2, 3, \ldots$ **do**
  for $local\_epoch \leftarrow 1, 2, 3, \ldots$ **do**
    for $step \leftarrow 1, 2, 3, \ldots$ **do**
      for $client \leftarrow 1, 2, 3, \ldots$ in parallel **do**
        $grad_{client} = Gradient(w_{client}, data_{client})$
        neighbors = GetRandomNeighbors(c, without = client.last)
        $w_{client} += Mean(neighbors.grad)$
        client.last = (client.last + neighbors)[-n:]

4.4.3 Heuristic 2: F1-score

The second and third heuristics utilize the models' performances to promote communication between clients with better-performing or dissimilar models. After each round, clients calculate their models' per-class F1-scores on a test set and share them with the network. Clients then select neighbors to communicate with based on the dissimilarity or similarity scores computed using these F1-scores.

$$\text{neighbor dissimilarity score} = ndc = \sum_{\text{class i}} \left| F_k^i - F_c^i \right| \tag{4.1}$$

FedAVG

Algorithm 4.7 FedAVG heuristic 2: F-1 Score

```
for round ← 1, 2, 3, ... do
    for client ← 1, 2, 3, ... do
        w_client = fit(w_client, data_client, epochs=$local_epoch)
    for client ← 1, 2, 3, ... in parallel do
        neighbors = GetNeighbors(c, without = client.last, metric = ndc)
        w_client ← Mean(neighbors.weight)
```

FedSGD

Algorithm 4.8 FedSGD heuristic 2: F-1 Score. c is fraction of clients that perform computation on each round

```
for round ← 1, 2, 3, ... do
    for local_epoch ← 1, 2, 3, ... do
        for step ← 1, 2, 3, ... do
            for client ← 1, 2, 3, ... in parallel do
                grad_client = Gradient(w_client, data_client)
                neighbors = GetNeighbors(c, without = client.last, metric = ndc)
                w_client+ = Mean(neighbors.grad)
```

4.4.4 Heuristic 3: Score cosine

As mentioned in Section 4.4.3, after calculating F1-Score per class, Clients select neighbors to communicate with based on the dissimilarity or similarity scores obtained using cosine score.

$$\text{neighbor dissimilarity score} = ndc = \cos(F_k, F_c) = \frac{F_k.F_c}{\|F_k\| \,.\, \|F_c\|} \tag{4.2}$$

FedAVG

Algorithm 4.9 FedAVG heuristic3: Score cosine. c is fraction of clients that perform computation on each round

```
for round ← 1, 2, 3, ... do
    for client ← 1, 2, 3, ... do
        w_client = fit(w_client, data_client, epochs=$local_epoch)
    for client ← 1, 2, 3, ... in parallel do
        neighbors = GetNeighbors(c, without = client.last, metric = ndc)
        w_client ← Mean(neighbors.weight)
```

FedSGD

Algorithm 4.10 FedSGD heuristic3: Score cosine. c is fraction of clients that perform computation on each round

```
for round ← 1, 2, 3, ... do
    for local_epoch ← 1, 2, 3, ... do
        for step ← 1, 2, 3, ... do
            for client ← 1, 2, 3, ... in parallel do
                grad_client = Gradient(w_client, data_client)
                neighbors = GetNeighbors(c, without = client.last, metric = ndc)
                w_client+ = Mean(neighbors.grad)
```

4.5 Experiments and Results

4.5.1 Experimental Setup

The experimental setup was conducted on a Windows 10 Pro operating system, running on a powerful hardware configuration comprising 64 GB of RAM and an Intel(R) Xeon(R) W-2155 CPU operating at 3.30 GHz. The system was further enhanced with an NVIDIA GeForce RTX 3060 GPU, boasting 12 GB of dedicated memory. To facilitate the experiments, the system was configured with CUDA version 11.7, Tensorflow 2.10.0, and Python 3.10.9.

4.5.2 Application of FL P2P for DFU classification and follow-up

The overall architecture we are proposing for classification of DFU is based on the Siamese network. The Siamese network was presented in the context of signature verification [124] and comprises of two identical networks that take in separate inputs, but are connected in the last layer.

Figure 4.3 gives a high level view of siamese networks as a block diagram. Siamese Neural Network usually use contrastive loss [252] which aim to maximize the proximity between positive pairs while simultaneously increasing the dissimilarity between negative pairs.

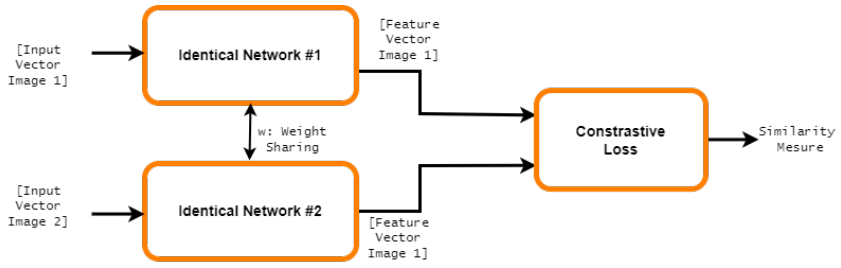


Figure 4.3: Block Diagram of Siamese Network.

For the CNN backbone we used EfficientNetV2S based on EfficientNet [152] architectures which have been shown to significantly outperform other networks in classification tasks while having fewer parameters. EfficientNetV2S has fewer parameters, making it more suitable for low-resource settings and it uses a combination of efficient network design and compound scaling to achieve high accuracy with fewer parameters [171]. The second backbone of the ensemble model is based on Vision Transformers. This was first introduced by the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [115] and is referred to as Vision Transformers (ViT).
The classification model is a milestone in the development of an innovative tool to be used to assist medical health professional in doing follow-up of patient having DFU. Figure 3.2 illustrates the proposed approach DFU classification.

4.5.3 Dataset

Data quality is a crucial factor that directly affects the performance of supervised learning algorithms. The utilization of a representative and high-quality dataset is critical for achieving optimal accuracy and performance [175]. In this study, we obtained the dataset from the DFU 2021 challenge organized by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) society [170]. The proper licensing was also secured for this research, ensuring that all ethical and legal requirements were met.

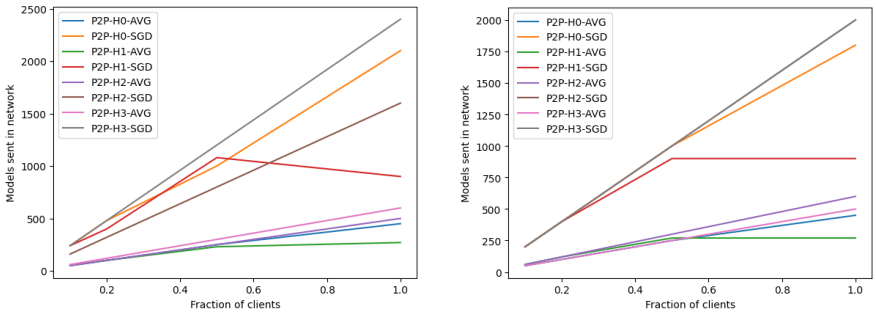
4.5.4 Experimental Parameters

We initially aimed to compare the performance of the centralized version of Federated Learning (FedAVG and FedSGD) with the distributed P2P architecture (FedAVGP2P and FedSGDP2P). The objective was to use a high number of clients (C=100, 200, 300, etc.) and a large number of communications (round=100, 200, 300, etc.) in our experiments to obtain the most relevant

results for the purpose of analysis. However, we soon realized that due to resource constraints, the computation times were excessively high, primarily because of the heavy deep learning models used and described earlier which we also had to substitute for a computation friendly backbone. As a result, we decided to limit the maximum number of clients to c=20 and the maximum number of rounds to 10. In the case of FedAVG, each round consists of two steps: selecting the clients that receive the aggregated model from the central server, and selecting the clients that send updates of their local models to the central server. In the case of FedAVGP2P and FedSGDP2P, during each communication round, we evaluate the clients' models on the test data. The round concludes when a client receives updates from all its neighbors. The training data are distributed among the clients, considering both IID and non-IID data distribution scenarios. To evaluate the performance of the three heuristics (n lastest, F1-score, and Score cosine), we vary the fraction of clients C with values of 0.1, 0.2, 0.5, and 1.0. As a result, each client communicates with 2, 4, 10, or 20 neighboring clients in each round. After each round, we assess all clients' performance on the test data. During experimentation the backbones initially proposed could not be used because of resource limitations. We were forced to change the backbones to a combination of ["MobileNet", "MobileNetV2"]

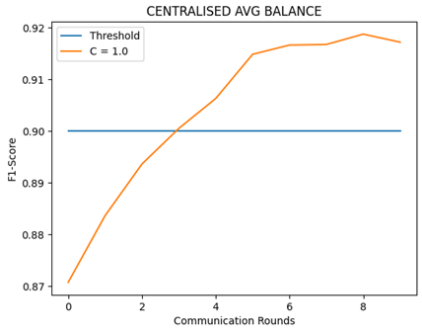
4.5.5 Results

In this section, we present the results obtained and examine the behavior of the centralized model (FedAVG) compared to the various decentralized FedP2P architecture variants, taking into account both IID and non-IID data distributions. An initial observation indicates that the FedP2P architecture, considering all the heuristics, appears to yield stable results compared to those obtained by the centralized FedAVG architecture. A detailed discussion of these results is provided in the discussion section.

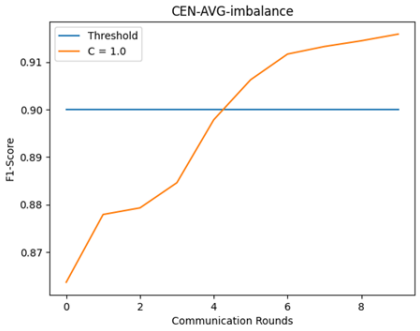


(a) IID **(b)** non-IID

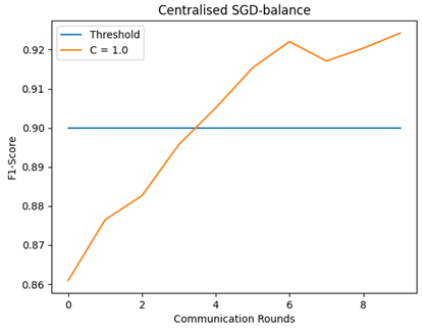
Figure 4.4: A comparison of FedAVG to FedAVGP2P considering models sent in the network when 90% model accuracy had been reached. C is the fraction of clients the central server (or every client with FedAVGP2P) had received updates from. According to the graph above, it can be said that Heuristic 1 has the best learning ability when we increase the factor C.



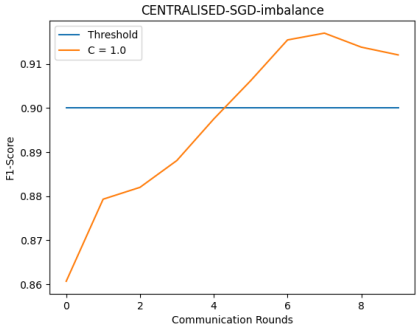
(a) Centralised FedAVG IID



(b) Centralised FedAVG non-IID

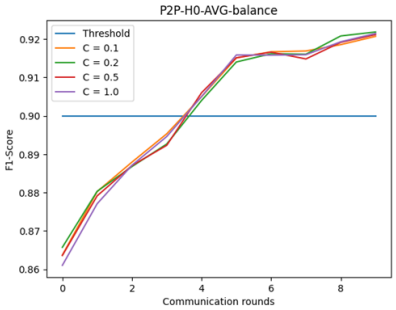


(c) Centralised FedSGD IID

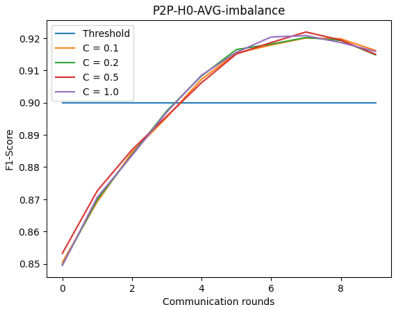


(d) Centralised FedSGD non-IID

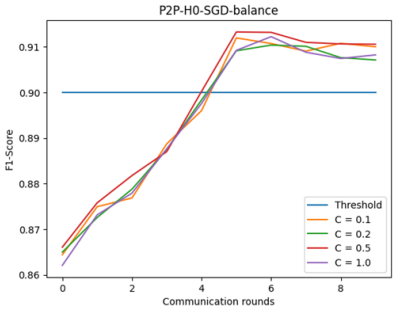
Figure 4.5: Centralised



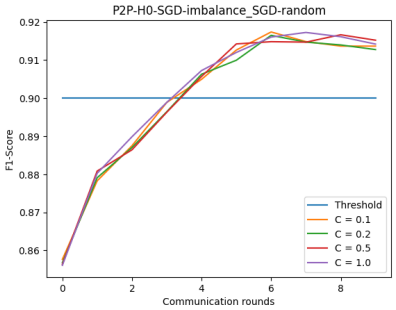
(a) P2P H0 FedAVG IID



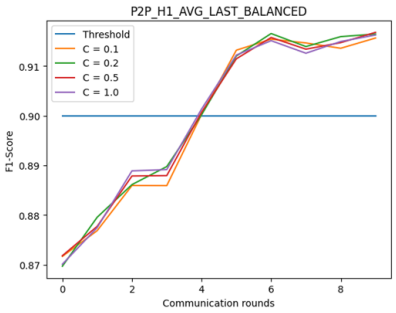
(b) P2P H0 FedAVG non-IID



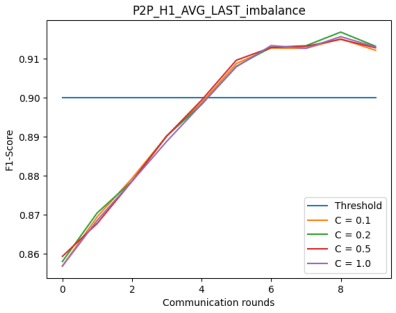
(c) P2P H0 FedSGD IID



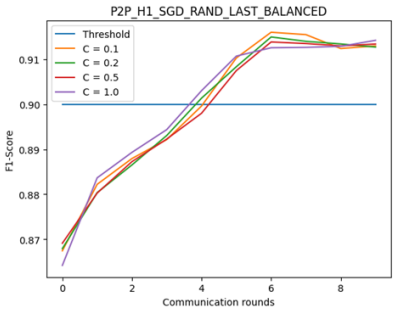
(d) P2P H0 FedSGD non-IID



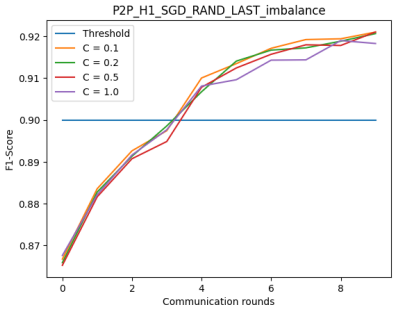
(a) P2P H1 FedAVG IID



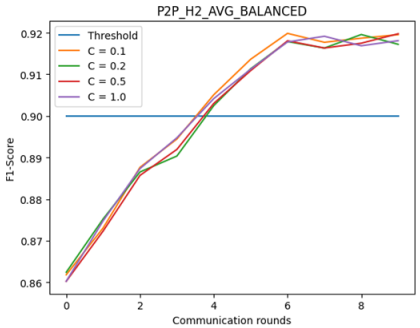
(b) P2P H1 FedAVG non-IID



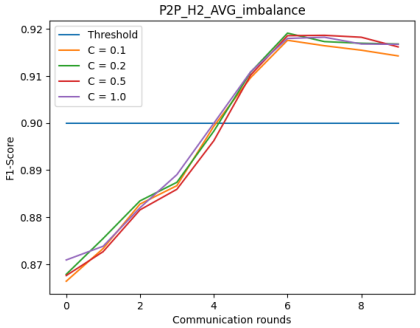
(c) P2P H1 FedSGD IID



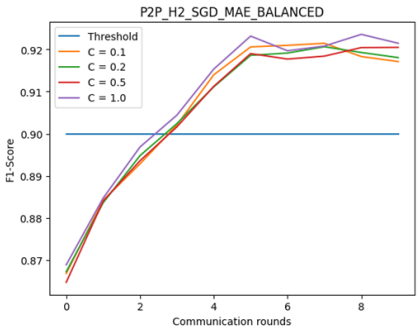
(d) P2P H1 FedSGD non-IID



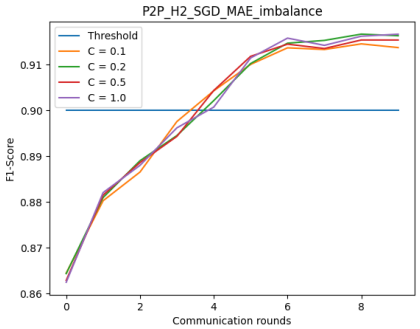
(a) P2P H2 FedAVG IID



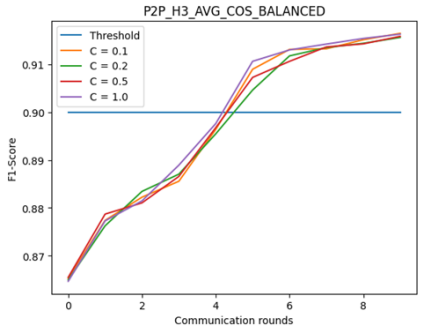
(b) P2P H2 FedAVG non-IID



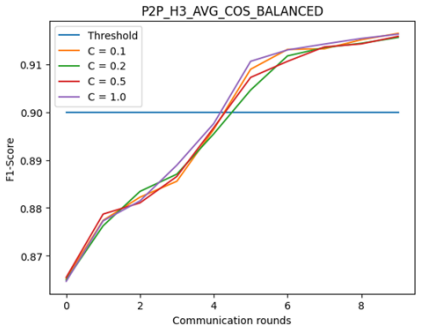
(c) P2P H2 FedSGD IID



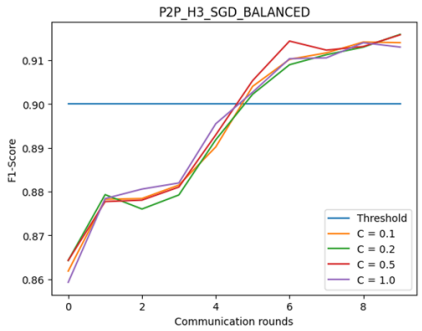
(d) P2P H2 FedSGD non-IID



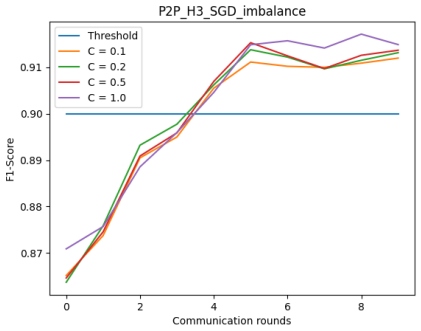
(a) P2P H3 FedAVG IID



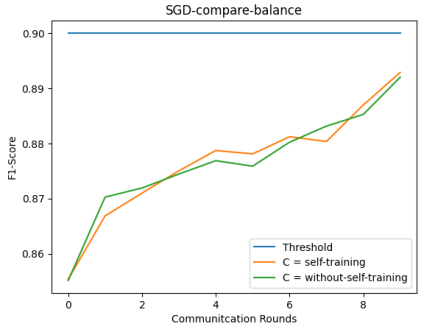
(b) P2P H3 FedAVG non-IID



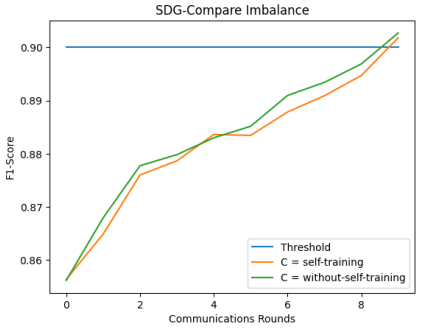
(c) P2P H3 FedSGD IID



(d) P2P H3 FedSGD non-IID



(a) IID



(b) non-IID

Figure 4.10: Compare a model that uses gradient vectors from its neighbors and both its gradient vectors (orange) and a model that uses only gradient vectors from its neighbors (green). Here we set the number of steps per round to 1.

4.5.6 Discussions

In this section, we discuss and analyze the results obtained. Considering the convergence behaviors of the models, the results indicate that the models trained with FedAVGP2P and FedSGDP2P can achieve comparable behaviors to Fedavg when provided with both IID and non-IID client data.

By observing the convergence behaviors of the models for FedAVG and FedP2P, we observe that the general behaviors are quite similar for both methods. Most experiments conclude with models reaching an accuracy of approximately 92%. These results suggest that the convergence behaviors of the average FedAVGP2P models are more comparable to FedAVG when the size of C is sufficient.
Let's consider the experiments with the fewest models sent over the network when the model accuracy reached 90%: in both cases, with IID and non-IID client data, both FedAVD and FedSGDP2P required higher network communication costs (number of rounds). However, naturally, with FedAVGP2P, the burden of communication costs is distributed among the participating clients rather than being heavily concentrated on a central server. Therefore, if there is a communication constraint at the central server level, such as insufficient bandwidth, FedAVGP2P may be a suitable approach.
Regarding the effect of the heuristics, for higher values of C, we observe comparable convergence behaviors for all the algorithms. This partly indicates that when communicating with a large portion of clients in the network, the choice of neighbors with whom each client communicates is not of significant importance. This situation makes us push our analysis further as to why the use of heuristics did not perform better than the original FedAVGP2P and FedSGDP2P. One possible reason could be that the heuristic leads the network clients to communicate more frequently with the same type of neighbors. This, in turn, could introduce multiple clusters in the network, where clients are more likely to communicate with neighbors within the same cluster. Additionally, this could prolong the time during which clients receive model parameters from neighbors outside their own cluster, potentially leading to lower performance by reducing the diversity of model parameters received by each client.
As future work, it would be interesting to investigate whether these clusters emerge by analyzing the choice of neighbors for each client throughout the training process. It would also be valuable to explore the scenarios in which FedAVGP2P or FedSGDP2P would be faster than FedAVG, taking into account the training time. The answer to this question depends on various factors, such as communication constraints and client systems.
For instance, using FedAVG could be a faster approach if the central server has sufficient bandwidth. However, FedAVGP2P could also be faster if the central server lacks such bandwidth. Looking at certain curves related to the heuristics of FedAVGP2P and FedSGDP2P algorithms, we observe the influence of accuracy achieved based on the number of rounds and the fraction of clients. This indicates the possibility of studying the trade-off between precision and communication according to the methods used.
Furthermore, at a fixed precision level, the different methods yield varying numbers of rounds, which can be utilized to measure the communication cost of each method. Similarly, we can explore and compare the methods to find the one that achieves the best precision at a fixed communication cost.

4.5.7 Limitations

A limitation of the study is the lack of resources to train richer models on our dataset of diabetic foot ulcers with our high performing deep learning Siamese model. It would be interesting to replicate the same simulation experiments by increasing the number of communicating clients in the case of FedAVGP2P and FedSGDP2P algorithms.

4.6 Conclusion

The overall results presented in this chapter indicate that training a model using a P2P FL architecture could be a viable approach for collaborative neural network modeling among multiple clients without sharing their training data. Firstly, the results show that models trained with FedAVGP2P and FedSGDP2P are comparable to models trained with the centralized FedAVG architecture in terms of accuracy. FedP2P may be less desirable due to higher global network costs compared to FedAVG, as more data need to be transmitted to achieve comparable model convergence behaviors. However, the use of a P2P topology offers several advantages, such as the absence of a single point of failure and dependence on a central server. This makes P2P FL a wise choice if these characteristics are required. To further refine the relevance of our results, additional measurements should be implemented by increasing the number of collaborative clients. Our results in P2P FL, through FedAVGP2P and FedSGDP2P, demonstrate it as a promising approach for training neural network models across multiple clients.

5

Chatbot for Diabetes

Summary

5.1 Introduction

Natural Language Processing (NLP) has emerged as a transformative technology at the intersection of linguistics and computer science. It is a multidisciplinary field focusing on giving machines the ability to understand, interpret, and generate human language in a way that is both meaningful and contextually relevant. There are several studies that have demonstrated how the associated techniques with NLP are being used in diverse fields. One of the most famous uses of NLP is ChatGPT from OpenAI and Bard from Google. NLP's impact extends to several sectors, such as finance, e-commerce, education, and entertainment, and NLP's impact extends to diverse sectors, including healthcare, finance, e-commerce, education, and entertainment. NLP provides valuable text and knowledge base-derived information for a variety of analytical purposes, including description, classification, data reduction, topic modeling, and sentiment analysis.
Figure 5.1 shows the Learn Nest with the objective of applying AI to the ecosystem in terms of adding a conversational agent which can provide round the clock assistance and hence enhance support, education and management for diabetes patients.

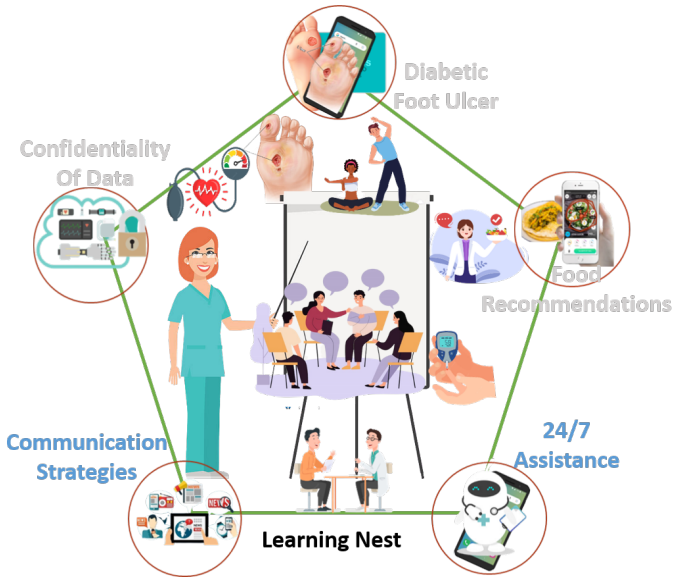


Figure 5.1: Application of Natural Learning Processing to AI powered ecosystem

5.1.1 Motivation and Research Question

Research in the field of conversational agents (CAs) encompasses numerous disciplines, including natural language processing, information retrieval, machine learning, and dialogue systems. Conversational agents, also known as chatbots, can communicate with humans and have become an active area of research, gaining popularity due to the emergence of artificial intelligence [253]. In the context of chronic diseases like diabetes, Large Language Models (LLMs) that provide daily recommendations can offer continuous guidance to patients, improving disease management and outcomes [254]. While there are various types of chatbots, our focus is on domain-specific medical chatbots, particularly those related to diabetes.
Intelligent conversational systems find applications across diverse domains such as customer support, education, e-commerce, healthcare, and entertainment. Numerous sub-tasks exist within the realm of conversational AI. A recent study on state-of-the-art CA systems categorizes them into three primary tasks: question answering, task-oriented dialogues, and social chatbots [255].

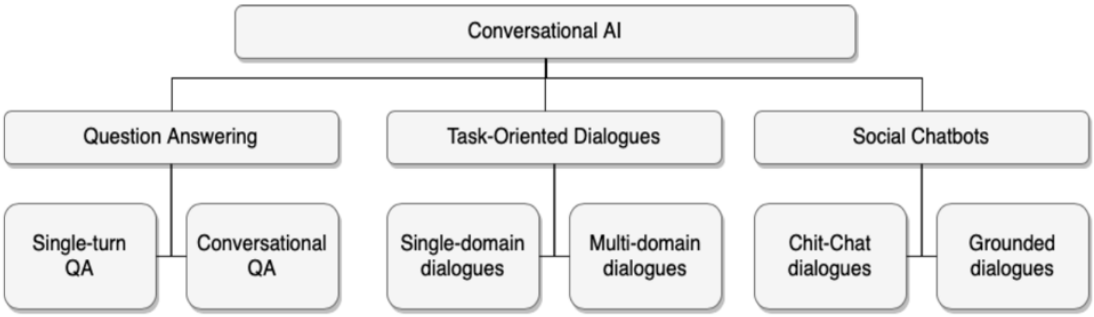


Figure 5.2: Tasks within conversational agent as identified by Gao *et al.* [255].

The Question Answering (QA) system is a crucial component in modern CA. It represents the intersection of NLP, ML, and Semantic Analysis [256]. QA can be perceived as a specialized form of one-way dialogue. While a QA system's primary objective is to provide accurate responses to natural language inquiries, dialogue systems emphasize producing contextually apt and coherent replies to user inputs. The recent progress in DL have paved the way for implementing QA through end-to-end generation. In this approach, a neural network centered around questions crafts responses, effectively catering to the nuances and diversities of human language. Significantly, this all-encompassing model is trained in a unified manner, eliminating the need for specialized linguistic knowledge, such as crafting a semantic analyzer. The emergence of QA systems marks a pivotal development in medical science. There has been an extensive exploration in the field of automated text-based QA agents, commonly referred to as chatbots, that utilize various neural networks [257].
The recent global pandemic has brought to forefront the growing significance of conversational agents in the healthcare industry. As the healthcare sector face overwhelming challenges,

these agents have proven to be invaluable in addressing a variety of health-related concerns. Recently, chatbots have garnered significant attention, especially with the introduction of the Chat Generative Pre-trained Transformer (CHATGPT) by OpenAI and Bard by Google. Chatbots serve as an excellent platform for delivering personalized education to patients, making them potent tools for managing chronic diseases [258], [259]. If a CA can dispense dependable information for minor issues—excluding direct treatment—it could afford healthcare professionals more time for patient care, potentially leading to cost savings [260], [261].
The research question addressed in this section concerns:

RQ-3: How to use deep learning models to implement chatbot capable of addressing inquiries related to diabetes?
The remainder of this chapter is structured as follows: Section 5.2 gives some background information on NLP. Section 5.3 provides an analysis of critical related work. Section 5.4 presents a detailed description of our proposed pipeline and components, which is part of the scope of this chapter, followed by a comprehensive presentation of experimental results in Section 5.5. We conclude with a summary of our findings and suggestions for future research in Section 5.6.

5.2 Background and Preliminaries

To construct efficient and resilient NLP systems, it is important to develop NLP pipelines and workflows tailored to accommodate diverse data types, tasks, and models.

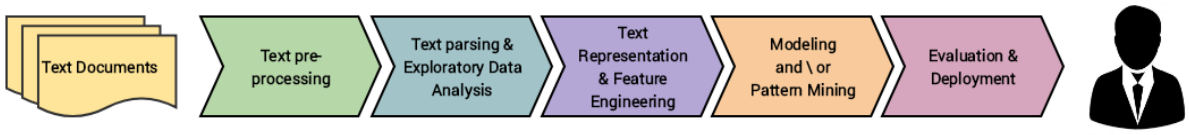


Figure 5.3: NLP project Workflow [262].

Figure 5.3 illustrates the standard workflow for implementing an NLP project. As with any machine learning project, the emphasis predominantly lies on the accessibility of data. The starting point involves a corpus of textual documents, which could be pre-existing or procured through web scraping. Another crucial phase encompasses pre-processing, which includes tasks such as parsing and elementary exploratory data analysis.

5.2.1 Text Pre-processing

Pre-processing text involved several activities which are explained below.

1. Tokenization: Consist of breaking breaks down text into smaller semantic units or single clauses. In tokenization, some stop words, such as "the", "a", will be removed as these words provide little useful information [263].

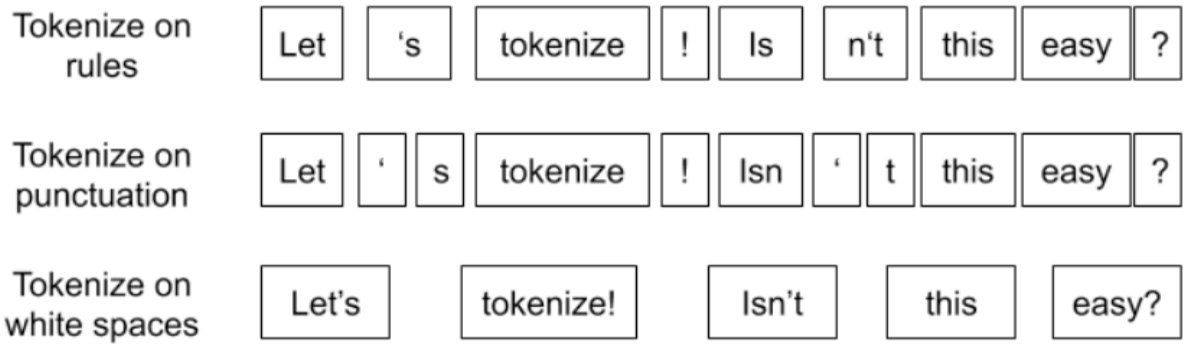


Figure 5.4: Example of application of Tokenisation to a sentence.

2. Stemming and lemmatization: This is a step of standardizing words by reducing them to their root forms

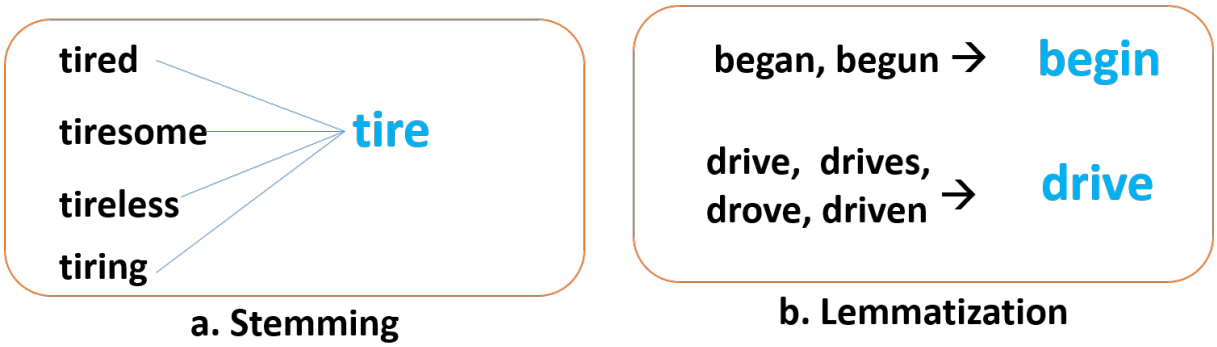


Figure 5.5: Example of a. Stemming and b. Lemmatization of text.

During the pre-processing phase, understanding the structure of the text is crucial. A mere assemblage of words without structure fails to convey any coherent meaning or provide substantive information. Knowledge about structure and syntax of language are mandatory. Part-of-Speech (POS) tagging and parsing are methodologies that assess lexical and syntactic attributes [263]. While POS tagging imparts lexical insights, parsing yields syntactic details. Parsing generates a tree mirroring the grammatical structure of a specific sentence, providing the relationships between its various components.

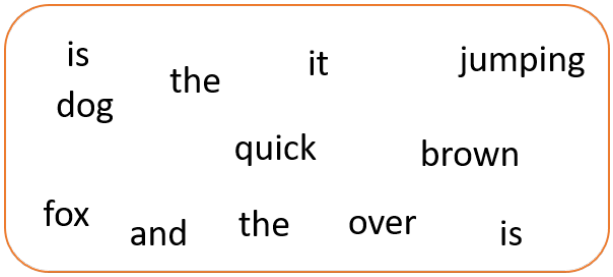


Figure 5.6: A group of unstructured word with no significant meaning. Inspired from [262].

1. Parts of Speech (POS) Tagging:

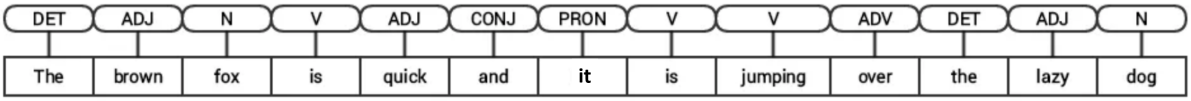


Figure 5.7: Illustration of POS tagging. Inspired from [262].

   - N: Nouns
   - V: Verb
   - ADJ: Adjective
   - ADV: Adverb
   - CONJ: Conjuction
   - DET: Determiner
2. Shallow Parsing: It is technique of analyzing the structure of a sentence to break it down into its smallest constituents (which are tokens such as words) and group them together into higher-level phrases. This includes POS tags as well as phrases from a sentence.

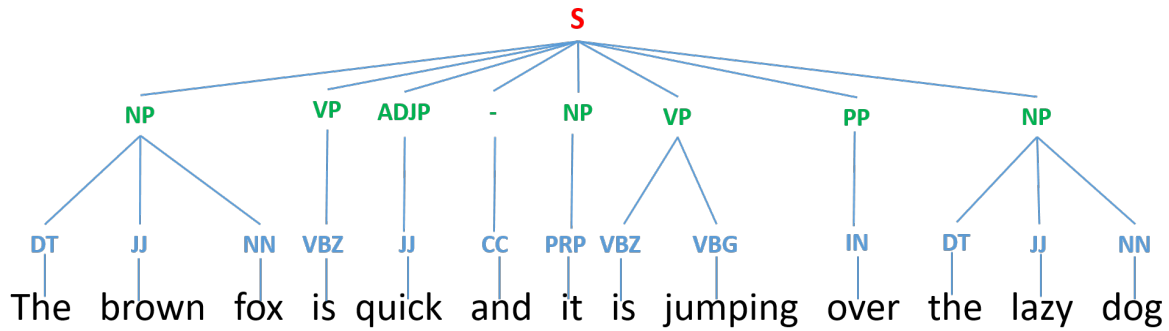


Figure 5.8: An example of shallow parsing showing higher level phrase annotations. Inspired from [262].

   - Noun phrase (NP): Phrases where a noun acts as the head word.

- Verb phrase (VP): These phrases are lexical units that have a verb acting as the head word.
- Adjective phrase (ADJP): These are phrases with an adjective as the head word.
- Adverb phrase (ADVP): These phrases act like adverbs since the adverb acts as the head word in the phrase.
- Prepositional phrase (PP): These phrases usually contain a preposition as the head word and other lexical components like nouns, pronouns, and so on.

3. Constituency Parsing: It is the process of analyzing the sentences by breaking down it into sub-phrases also known as constituents.

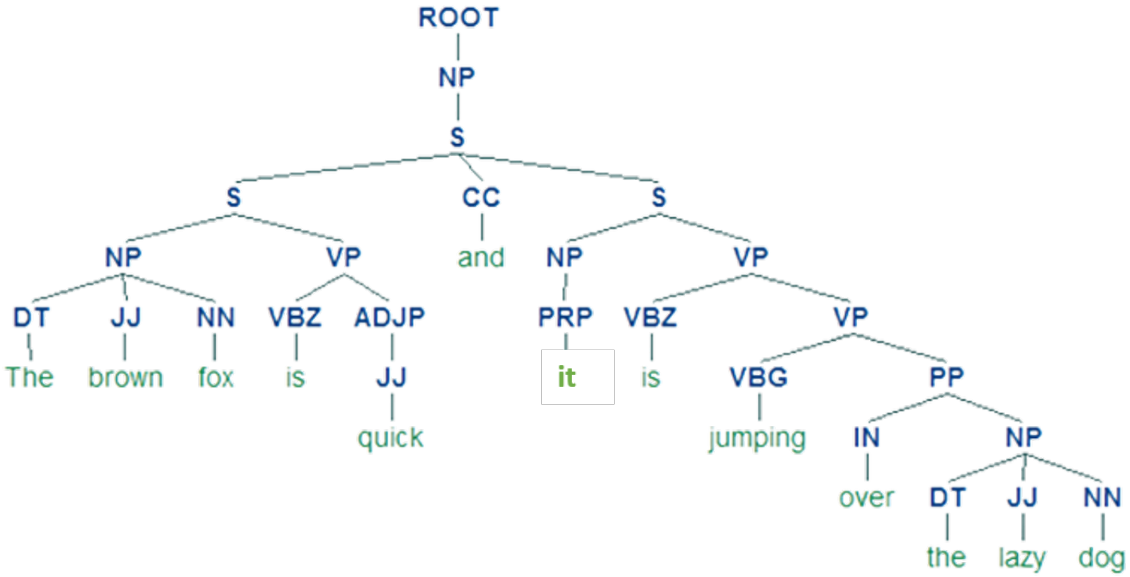


Figure 5.9: An example of Constituency parsing [264].

4. Dependency Parsing: It consists of analyzing the grammatical structure of a sentence by establishing relationships between "head" words and the words which modify those heads.

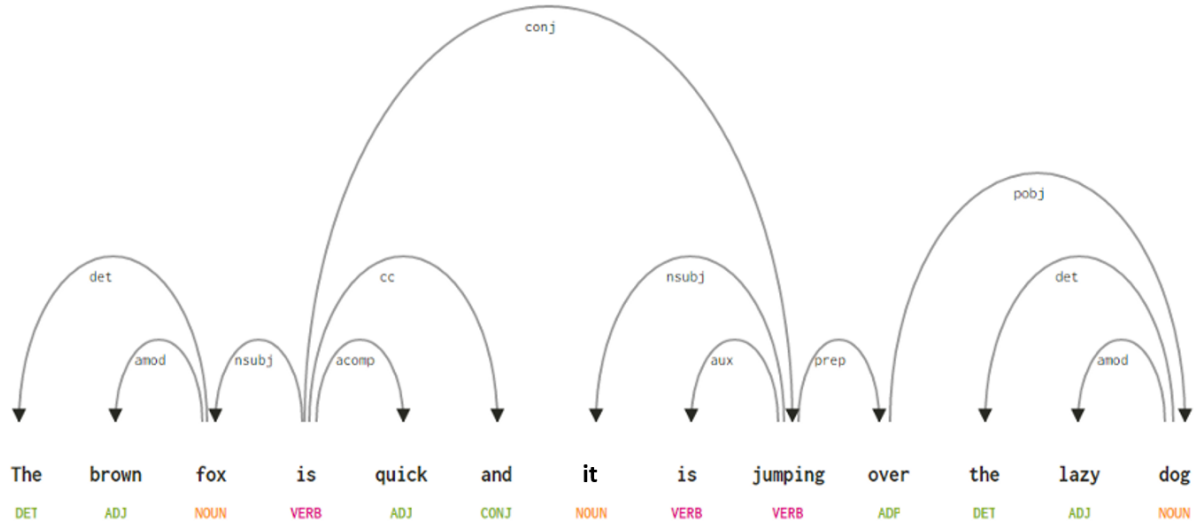


Figure 5.10: An example of Dependency parsing [264].

5.2.2 Bag of Words (BoW)

BoW is a technique for extracting features from text for in modeling when dealing with machine learning algorithms. It provides a textual representation that quantifies the frequency of words within a specific document. This technique relies on a predefined list of known words and assesses their presence, as illustrated in Figure 5.11.

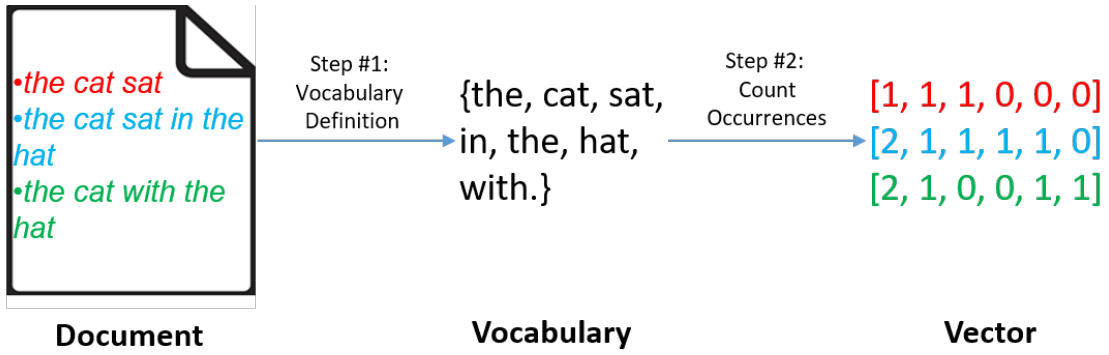


Figure 5.11: Demonstration of BoW.

BoW gives information on what words occur in the document and discard where exactly they occurred.

5.2.3 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is used for extracting information from text documents. It works on the idea that the importance of a word in a document

is determined by two factors:

1. how frequently the word appears in the document (Term Frequency or TF)
2. how unique or rare the word is across the entire corpus of documents (Inverse Document Frequency or IDF).

5.2.4 Word2Vec

Word2Vec [112], introduced in Section 2.4.7, is a technique for generating word embeddings. It creates word embeddings using two methods, both based on neural networks: Skip Gram and Continuous Bag Of Words (CBOW).
It combines the two factors, TF-IDF and gives a weight to each word in a document, with higher weights given to words that are both frequent in the document and rare in the corpus.

5.3 Related Work

Reddy *et al.* [265] focused on enabling machines to answer questions based on the conversation. They introduced a new dataset of conversational question-answering systems (COQA). They obtained an F1-score of 65.4% compared to human performance, which was 88.8%. The dataset used included 127000 question-answer pairs from 8000 conversation passages across 7 distinct domains. One interesting part of this work is the rationale that is included when a question is answered based on a passage. The creation of the dataset involved using human annotators. The model they try to implement is given a passage p, the conversation history $\{q_1, a_1, \ldots q_{i-1}, a_{i-1}\}$ and a question q_i, the task is to predict the answer a_i. Gold answers $a_1, a_2, \ldots, a_{i-1}$ are used to predict a_i. They use LSTMs [266], GloVe [267] and fastText [268] during experimentation. Chan *et al.* [269] use a pre-trained BERT language model to tackle question generation tasks. They experimented with three models, which are BERT directly and another two models by restructuring the use of BERT. The models proposed are evaluated on SQuAD dataset. They cite results obtained from improving state-of-the-art performance, which advances the BLEU 4 score of the existing best models from 16.85 to 22.17. The pre-trained BERT model used is $\text{BERT}_{\text{base}}$ and for sequence decoding, the algorithm used is Beam Search Strategy with beam size set to 3. Neural Generative Question Answering (GENQA) is proposed by Yin *et al.* [270]. GENQA can generate answers to simple factoid questions based on the facts in a knowledge base. The model is trained on a dataset composed of real-world question-answer pairs. The model is trained on a dataset composed of real-world question-answer pairs associated with triples in the knowledge base. Zhang *et al.* [271] proposed an ensemble technique to combine outputs from several deep learning models to achieve an F1-score of 77.96% for QA tasks. They found that a convolutional neural network (CNN) built on top of a BERT model achieved an F1-score of 76.56%. Using a BiLSTM encoder with bi-directional attention they yielded

an F1-score of 76.37%. The authors also pointed out that the pre-trained BERT architecture with an encoder-decoder on top is effective, as is adding BiLSTM or GRUs architectures.

Table 5.1: Chatbot /Question Answering in Medical Domain

Research Work	Year	Domain	Main Focus	Dataset	Techniques / Model	Metrics / Result
Zhang *et al.* [272]	2018	Medical	Integration of information from multiple text documents that are retrieved from a large-scale database as being most semantically relevant to a question-answer pair	Medical reports and examination reports of patients, medical textbooks and articles	Word embedding LSTM-based network with dual-path attention	Relevant document ratio: 0.29 Accuracy: 74.4 %
Zhou *et al.* [273]	2019	Medical	A BioBERT transformer model is proposed to extract semantic relation between the question and the answer	MEDIQA2019-Task3-QA	Transformer based seq2seq model	Accuracy of 76.24% Spearman of 17.12%
Bao *et al.* [274]	2020	Medical	Using different models like hierarchical Bi-Directional LSTM with attention, Siamese LSTM, BERT to calculate the semantic similarity between user query and questions in the database	Knowledge graph constructed from medical data collected from the Internet	LSTM-based models is used to compute a semantic similarity score for retrieving the answer	Average Evaluation Accuracy Average predict accuracy
Harilal *et al.* [275]	2020	Medical: Depression	Medical advice and empathetic response facilitated through a binary intent classifier in the initial stage	Empathetic dialogues datasets and Medical question answer dataset	LSTM-based models trained separately on two different datasets	The BLEU score = 0.179, The BERT score = 0.83. Accuracy of intent classifier = 98.5% Emotion classifier= 92.4%.
Soni *et al.* [276]	2020	Open Domain and Clinical and Biomedical	Evaluate the performance of various Transformer language models	CliCR and emrQA Fine tuing on SQuAD	BERT / BIOBERT, Clinical BERT	Exact Match=70.56
Alzubi *et al.* [277]	2021	Medical: Covid19	Document retriever using term frequency count followed by finding suitable answers in the retrieved documents using Distil-BERT transformer	Covid-19 QA	TF-IDF BERT DistilBERT	Exact Match=80.6 % F1-score=87.3 %
Yadav *et al.* [278]	2022	Medical	Presents a question-driven text summarization of the clinical text as the answer to the user query	MeQSum	Beam search algorithm with a beam size of 5 LSTM and other Transformers	ROUGE-L (F1) score =47.8%

Most of the work studied above uses either gold standard questions or answers for comparison of performance. We are aiming to propose a pipeline that generates questions and answers without human intervention. Though we want to demonstrate our framework for the close domain of diabetes, our aim is to have a general-purpose framework that can be adapted to any domain by changing the context dataset.

5.4 Proposed Solution

The chatbot pipeline plays a critical role in processing user queries. Upon receiving an input query from a user - whether they are pre-diabetic, diabetic, or simply seeking relevant information - the pipeline understands the context and intent of the query, subsequently generating a contextually appropriate response. To enable this functionality, it is imperative to train the chatbot model using a dataset consisting of context, questions, and corresponding answers. In the field of diabetes-related queries, the availability of a rich and pertinent text corpus is crucial. This corpus serves as the foundation for generating context, formulating relevant questions, and providing accurate answers. The subsequent sub-sections explain the various sub-systems that constitute the chatbot pipeline, detailing their functionalities and contributions to the overall system. Figure 5.12 shows the overall architecture of the chatbot pipeline.

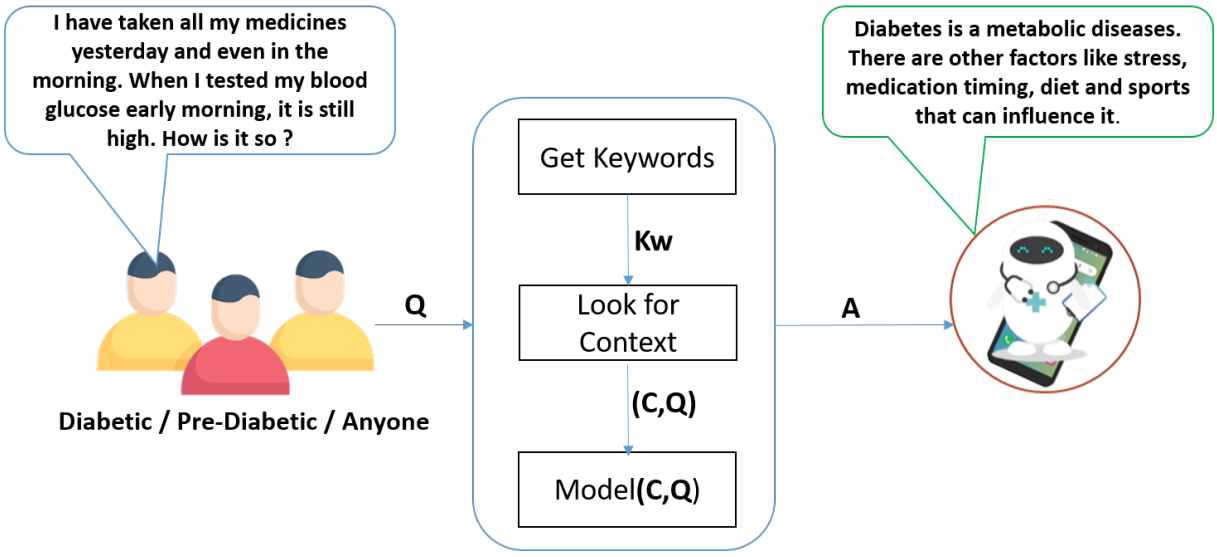


Figure 5.12: The proposed chatbot Pipeline.

5.4.1 Question/Answering Pipeline

The question-answering framework is central to the system's architecture. Figure 5.13 illustrates the complete pipeline, encompassing stages from data acquisition to the formulation of the dataset tailored for training the conversational agent.

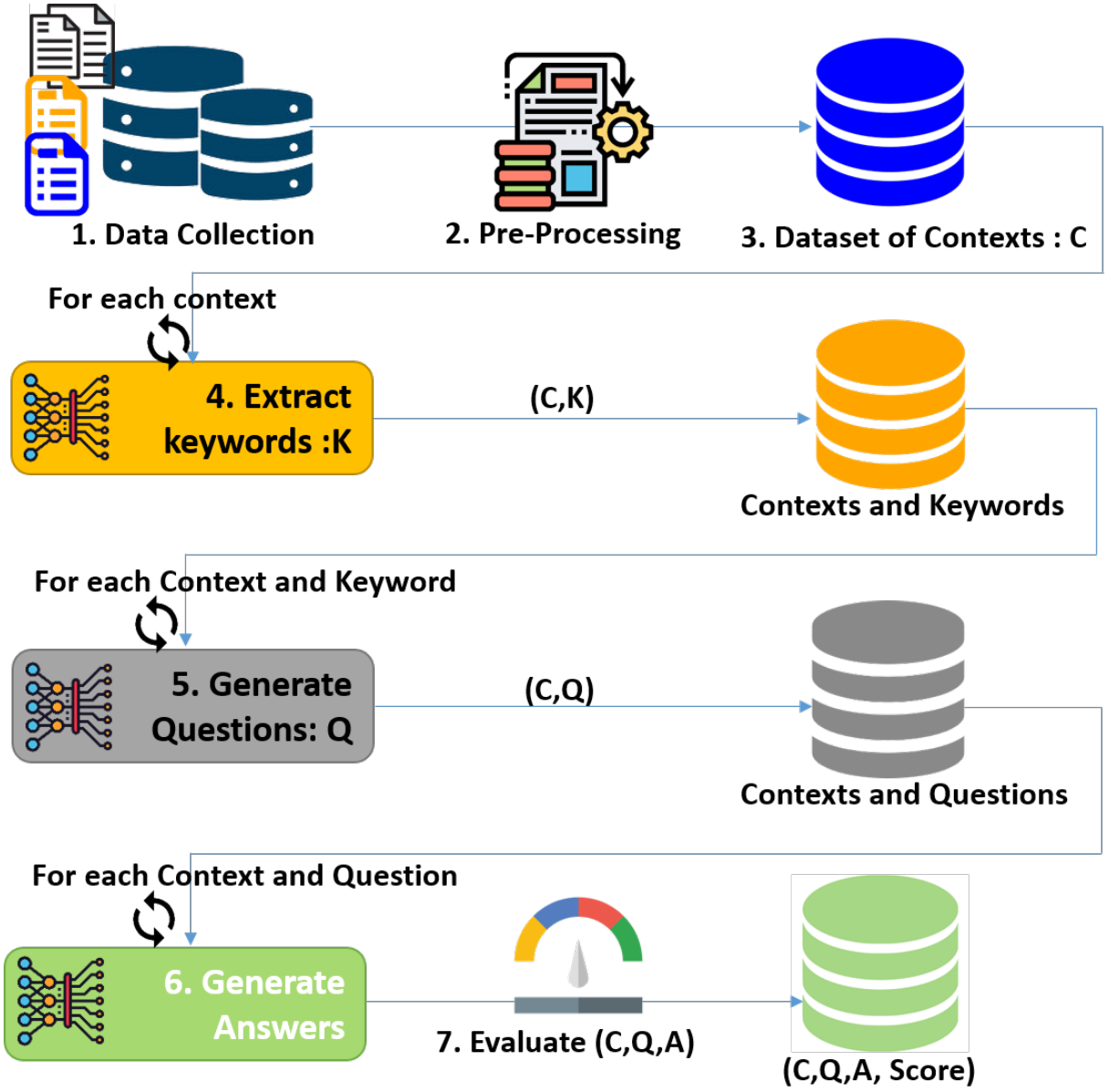


Figure 5.13: The proposed Generic Question Generation and Answering Pipeline.

The Question Generation and Answering pipeline works as follows:

1. A corpus of data on a specific subject has to be collected and pre-processed. The corpus of data can be from diverse but trusted sources. In this case, the closed domain being targeted will be diabetes-related.
2. We then direct our focus to a very critical phase within the NLP workflow: data cleaning and pre-processing. This integral process is introduced in Section 5.2 and encompasses a range of essential tasks, including but not limited to Text Tokenization, Stopword Removal, Stemming and Lemmatization, and Part-of-Speech Tagging, among others.

3. Once the previous step is completed, we are left with a meticulously cleaned and pre-processed text corpus featuring diabetes-related information, which is the context that will be exploited for the question-answering part.
4. For each piece of content in the corpus, keywords need to be extracted with the aim of generating the most questions. In this study, the aim is not to answer only the WHAT question, but to also target WHY, HOW, and WHEN. Figure 5.14 provides an illustration of a context with some keywords highlighted that can be used as the basis for the generation of questions.

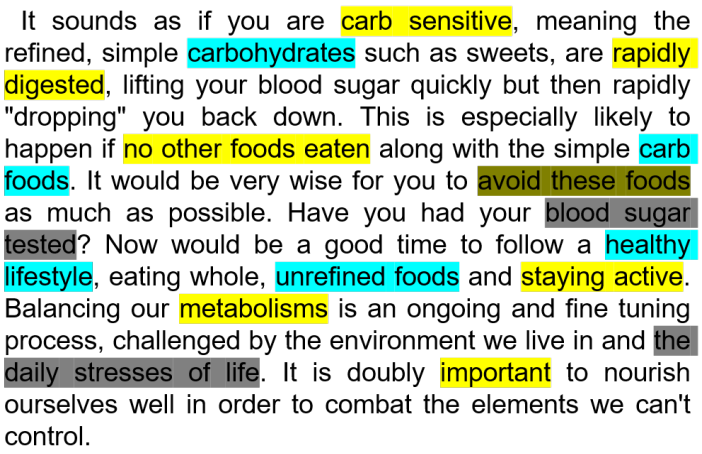


Figure 5.14: Illustration of a Context with highlighted Keywords and key phrases.

5. Using the context (C) and keyword (K), we can now proceed to generate a comprehensive list of questions. This process can be accomplished through the utilization of pre-trained deep learning models. To assess the effectiveness of these models, it is imperative to subject them to rigorous testing, evaluating the quality and diversity of questions they generate when provided with a specific context and keyword.
6. Once both the context and questions are available, we possess the required inputs to engage another deep learning model. This second model is responsible for identifying answers within the context, specifically addressing the questions associated with it.
7. As the final step, as with any deep learning task, it is crucial to evaluate the ultimate outcome. The selection of an appropriate evaluation metric must be carefully considered and tailored to align with the specific requirements of the task at hand. The main metrics to be considered will be F1-score, BERT score, and RQUGE, which will be dealt with in more detail in Section 5.5

5.4.2 Keyword Extraction

The domain of keyword extraction has been a dynamic field of research for an extended period, encompassing a wide array of applications in Text Mining, Information Retrieval, and Natural Language Processing [279]. Keyword extraction can be done using supervised, semi-supervised, and unsupervised algorithms. According to findings in literature, the pre-trained model KeyBERT exhibits good performance for keyword extraction. Our aim with keyword extraction in a context is to identify a group of important words or phrases that can reflect the main ideas, as shown in Figure 5.15. It is well documented that when dealing with textual data, context is very important, as we need to catch semantic meaning. Literature reviews suggest that the pre-trained model, KeyBERT [280], stands out in its performance for keyword extraction tasks. Our objective in contextual keyword extraction is to discern significant words or phrases that encapsulate the central themes, as depicted in Figure 5.15. For keyword extraction, we propose to test performance with KeyBERT and Keyphrase Boundary Infilling with Replacement (KBIR) [281] fine-tuned on the INSPEC dataset, which is commonly called KBIR/INSPEC. The INSPEC dataset is a collection of scientific papers from the fields of computers, control, and information technology. It was created by the Institution of Engineering and Technology (IET) and contains over 2000 papers, each of which is annotated with key phrases.

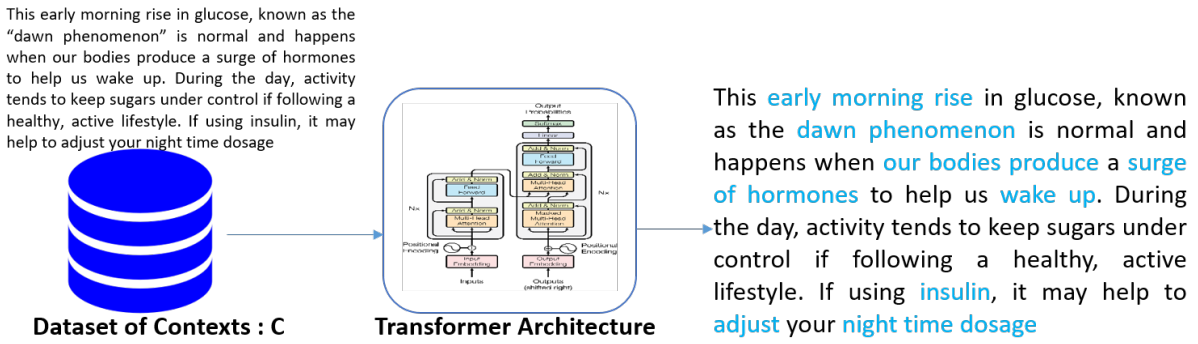


Figure 5.15: Block diagram of proposed Keywords/Phrases extraction using Transformer based Architecture.

KeyBERT [280] is a keyword extraction technique that uses BERT embeddings to identify the most representative keywords within a given text document. This unsupervised method comprises three sequential steps: candidate keywords or keyphrases, BERT embedding, and similarity measurement, as shown in Figure 5.16.

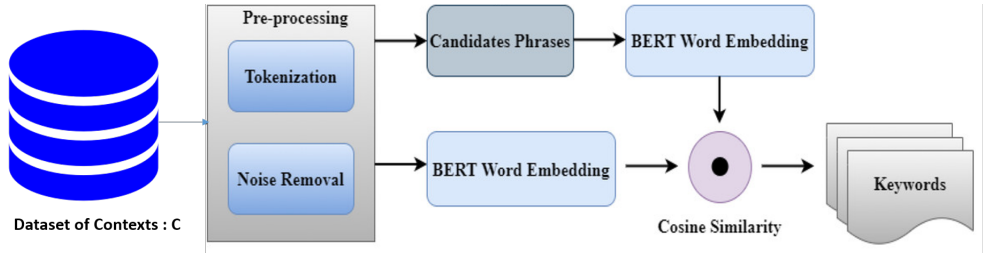


Figure 5.16: Architecture of the KeyBERT model for keywords extraction using BERT Embeddings inspired from [282].

KBIR is a pre-trained model for keyphrase extraction. It is a transformer model that is fine-tuned on a multi-task learning setup for optimizing a combined loss of Masked Language Modeling (MLM), Keyphrase Boundary Infilling (KBI), and Keyphrase Replacement Classification (KRC).

5.4.3 Question Generation

Question generation (QG) is the task of automatically creating questions that can be answered by a certain span of text within a given passage, is important for question-answering [283]. This task accepts context and keywords as inputs and generates questions that emphasize the specified keyword. QG it is a challenging problem in AI, which aims to generate natural and relevant questions from natural language text [284].

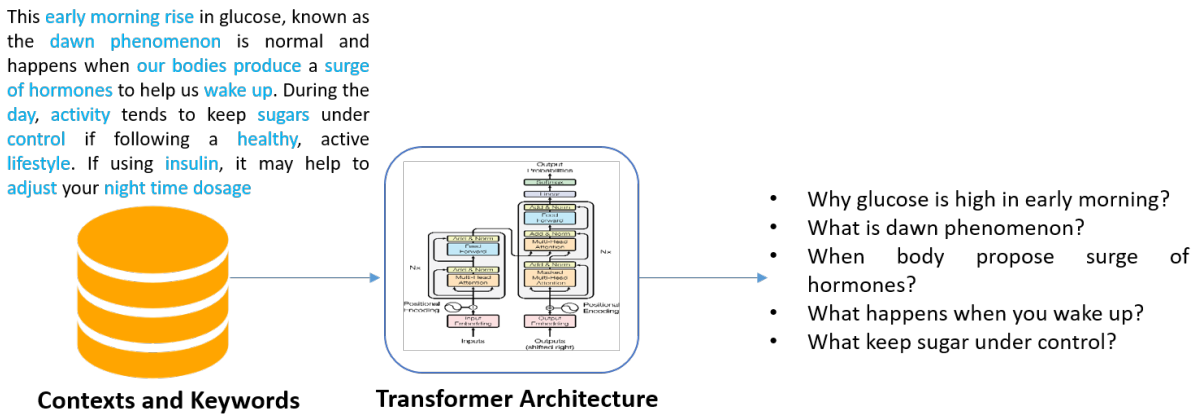


Figure 5.17: Block diagram of question generation using Transformer based architecture.

We propose to use T5, Text-to-Text Transfer Transformer [285], which is pretrained on a large corpus of text data using a denoising autoencoder objective. After pretraining, it can be fine-tuned on task-specific data with task-specific objectives. T5 has achieved state-of-the-art performance on a wide range of NLP benchmarks and competitions, demonstrating its effectiveness and versatility across different tasks.

5.4.4 Answers Generation

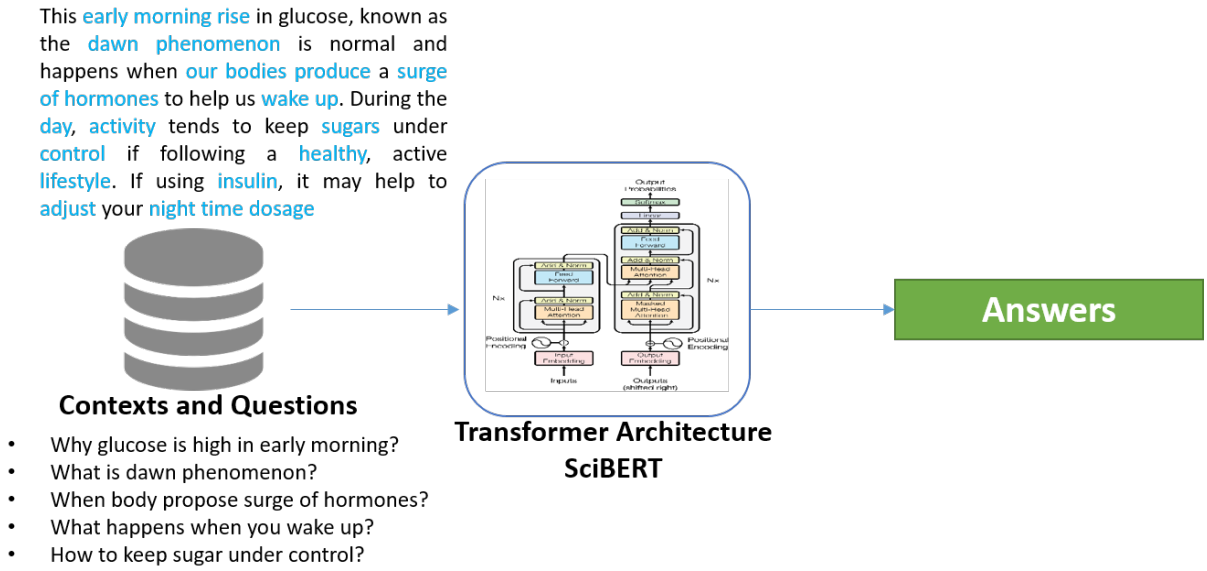


Figure 5.18: Proposed Answer generation using pre-trained model Sci-BERT.

SciBERT [286] is a pretrained language model based on BERT and has been shown to achieve state-of-the-art results on a variety of question answering benchmarks, including SQuAD and QuAC. It is particularly effective at answering questions about complex scientific topics. It is specifically pre-trained on scientific language. To use SciBERT for question answering, you can simply input the question and the context passage into the model, and it will output the answer. SciBERT can be used through the Hugging Face Transformers library or through a variety of online tools and APIs. Figure 5.19 shows how the Sci-BERT model was trained.

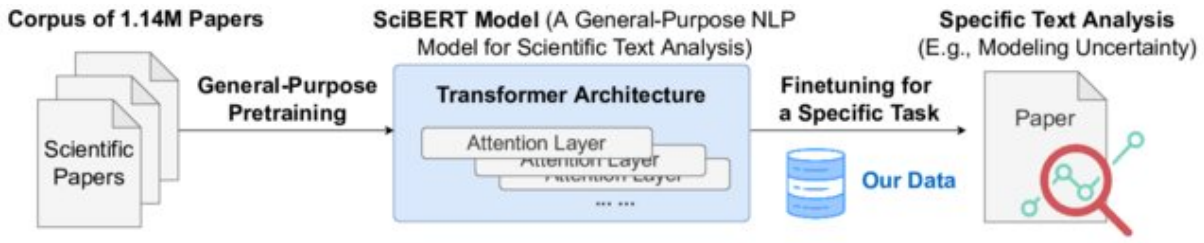


Figure 5.19: Block diagram od Sci-BERT a pre-trained language model [287].

5.4.5 Chatbot Training

A question-answering model is not sufficient for functioning as a relevant chatbot. In a chatbot, a user simply inputs a question and expects a response. However, in our question-answering model, we require both a question and context in order to generate a relevant response. We

design a chatbot pipeline that ensures the user inputs a question and retrieves a mini-context in the form of textual data that corresponds to the question asked, in connection with a global context of the pipeline defined by a large corpus of unannotated textual data. Using the generated mini-context and the question, the Chatbot pipeline generates the most appropriate response to the question asked. Features of the previous component, such as keyword extraction and question generation (QA), have been used to enhance the quality of the generated response. Figure 5.20 illustrates the training of the chatbot.

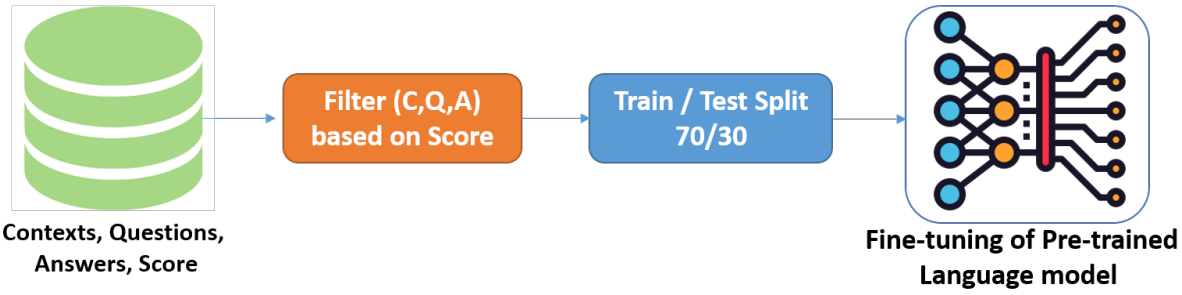


Figure 5.20: Block diagram of the chatbot Training.

The large language model we are proposing to fine-tune with our context, questions, and answers dataset will depend on which one exhibits the best performance. The models we are going to test on a snapshot of the dataset are:

1. PubMedBERT [288] a model that is tailored for biomedical and clinical natural language processing (NLP) tasks, particularly for tasks involving the analysis of scientific literature and biomedical text. It is pre-trained on a vast corpus of text from PubMed, a widely used repository of biomedical and life sciences literature, which includes articles, abstracts, and other scientific publications.
2. BioBERT [289] is a model designed specifically for biomedical and clinical natural language processing (NLP) tasks. While the original BERT model is pre-trained on a diverse range of text from the internet, BioBERT is pre-trained on a large corpus of biomedical and clinical text, which includes scientific articles, electronic health records (EHRs), medical literature, and other healthcare-related documents.
3. DistilBERT [290] is a variant of the BERT model that has been distilled or compressed to make it smaller and faster while retaining much of its language understanding capabilities.
4. BioLinkBert [291] BioLinkBERT-base model is pretrained on PubMed abstracts along with citation link information. LinkBERT is a transformer encoder (BERT-like) model pretrained on a large corpus of documents. It is an improvement of BERT that newly captures document links such as hyperlinks and citation links to include knowledge that spans across multiple documents.

5.5 Experimentation and Results

5.5.1 Dataset

In any machine learning workflow, the importance of a well-structured and comprehensive dataset cannot be overstated. Specifically, in the development of conversational agents tailored for addressing healthcare concerns, such as diabetes in this current study, a rich and domain-specific corpus of textual data is mandatory. The initial dataset, which constitutes the starting context, was collected from the following sources:

1. World Health Organist ion
2. American Diabetes Association
3. Online media articles on Diabetes

Web scraping methods, including Beautiful Soup, ParseHub [292], and Octoparse [293], were employed to extract electronic text data from structured web pages of news outlets and health magazines with a specific focus on diabetes. Subsequently, the gathered data underwent consolidation and filtration, with the criteria being articles published within the last five years to ensure their pertinence to contemporary lifestyle practices and prevailing diabetes conditions. Following this, pre-processing techniques were implemented to render the textual data ready for analysis.

It is worth highlighting that the data collection process opened up the possibility of performing survey-like analysis on the data extracted from online media articles. The findings of this survey were subsequently published in a conference paper titled **Analyzing Online Media Articles on Diabetes using Natural Language Processing: A Comparative Study of the Indian Ocean Region and France** available in Annex A.1.

The data collected were preprocessed and cleaned into a final list of 179 contexts, as shown in Figure 5.21.

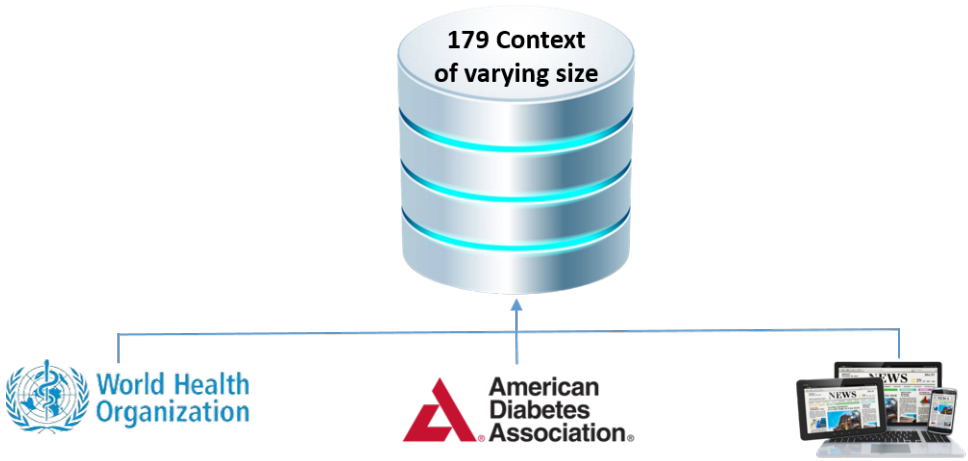


Figure 5.21: Results of Dataset collection from diverse sources.

Data Augmentation

During data collection process the context were found to be of varying size and this also limited the number of context to only 179. We applied data augmentation techniques in the following forms:

1. Sentence Split: Splitting contexts in sentences
2. CC Split: Splitting contexts on Coordinating Conjunctions." Coordinating conjunctions are words like "and," "but," "or," "nor," "for," "so," and "yet" that are used to connect words, phrases, or clauses of equal grammatical rank in a sentence. When splitting contexts on CC, it means that you are dividing or segmenting a larger text or sentence into smaller parts at the locations where coordinating conjunctions (CC) are found.
3. IN Split: Splitting contexts into smaller parts at locations where prepositions (IN), which are words like "in", "on", "at" and "by" are found. When splitting contexts on IN, you are essentially identifying prepositions as indicators of relationships or locations within a text and using them as natural points to break the text into meaningful segments.
4. WRB Split: Splitting contexts into smaller parts at locations where wh-adverbs or wh-relative adverbs (WRB) are found. Wh-adverbs include words like "when," "where," "why," and "how," which are used to introduce questions or clauses that seek specific information about time, place, reason, or manner.
   We used the library of NLTK (Natural Language Toolkit) which is an open-source Python library and platform for working with NLP. One of the essential features of NLTK is its support for Part-of-Speech (POS) tagging.

5.5.2 Experimental Setup

The experiments were conducted several computer systems and two of the most powerful are mentioned below:

1. The configuration of the first system operates on a Windows 10 Pro operating system running on a powerful hardware configuration comprising 64 GB of RAM and an Intel(R) Xeon(R) W-2155 CPU operating at 3.30 GHz. The system was further enhanced with an NVIDIA GeForce RTX 3060 GPU, boasting 12 GB of dedicated memory. To facilitate the experiments, the system was configured with CUDA version 11.7, Tensorflow 2.10.0, and Python 3.10.9
2. The configuration of the second system operates on a Windows 10 operating system running on a powerful hardware configuration comprising 32 GB of RAM and an AMD Ryzen 5 3600 (3.6:4.2GHz) CPU and GPU. The system was further enhanced with an AMD Ryzen 5 3600 operating at 4.2 GHz. The system was configured with Tensorflow 2.11 and CUDA 11.2. To interface with huggingface transformers 4.28 Torch 1.12.1 and CUDA 11.3 were used.

5.5.3 Results

The primary aim of the question generation module was to produce a significant volume of questions while attaining a performance level comparable to that of human-generated questions. The primary objective of our question generation module was to produce a large number of questions, while ensuring the quality was comparable to that of human-generated questions. To evaluate the model's performance, it would necessitate human intervention to create a reference set of questions for comparison with those generated by our model. This would, however, contradict the goal of automated question generation. We have advanced in a systematic manner to generate an extensive dataset encompassing context, questions, and answers for training the proposed chatbot. The following section elaborates on the steps involved and presents the corresponding results.

Keyword Generation

In the pipeline we proposed, given the context, KeyBERT was planned to be used for keyword extraction. KeyBERT, as previously explained, is designed to provide an efficient and effective way to extract key terms from textual data. KeyBERT uses cosine similarity to calculate the similarity between the document embedding and each of the candidate keyphrase embeddings. Cosine similarity is a measure of the angle between two vectors. A cosine similarity score of 1 indicates that the two vectors are perfectly aligned, while a cosine similarity score of 0 indicates that the two vectors are orthogonal. The candidate keyphrases with the highest cosine similarity scores are then selected as the keywords for the document. While this method is simple and effective during execution on our dataset, it took a very long time to execute due to the generation of candidate keywords or keyphrases.
We opted to switch to an alternative method, specifically the KBIR/INSPEC model. KBIR is tailor-made for extracting keyphrases from scientific papers. It employs a pre-trained transformer model to pinpoint sections of the document likely to contain keyphrases. After determining the keyphrase boundaries, KBIR utilizes a replacement algorithm to extract these keyphrases. This algorithm functions by iteratively substituting the keyphrase boundaries with tokens deemed most likely to be keyphrases. This iteration continues until the algorithm settles on a set of keyphrases that are believed to be most pertinent to the document. Our initial approach aimed to generate a single question for each keyword or keyphrase. The outcomes of this procedure are detailed in Table 5.2

Table 5.2: Keyword Generation with Single Question Generation strategy

Dataset	Model	Num Keywords
Initial	KeyBERT	895
Initial	KBIR	1948
Sentence Split	KBIR	3460
CC Split	KBIR	2810
IN Split	KBIR	**4273**
WRB Split	KBIR	4066

Different strategies for splitting or processing the data (e.g., Sentence Split, CC Split, IN Split, WRB Split) have a noticeable effect on the number of keywords generated, with the "IN Split" yielding the highest number of keywords using the KBIR model.
Another strategy was adopted which is Multiple Question Generation from keywords and results are shown in Table 5.3

Table 5.3: Keyword Generation with Multiple Question Generation

Dataset	Model	Num Keywords
Initial SQG	KBIR	1948
Initial MQG	KBIR	15584
Sentence Split	KBIR	28993
CC Split	KBIR	**111666**
WRB Split	KBIR	93077

Question Generation

For question generation, we initially experimented with BERT and ProphetNet [294]. ProphetNet is a language model developed by Microsoft AI and is based on a pre-trained transformer architecture. It has the capability to generate questions pertinent to a specified set of keywords. However, during our experiments, all questions (100%) generated by ProphetNet began with the interrogative word "WHAT". Consequently, we explored alternative models.
Based on the literature review, we opted to experiment with the T5 model. T5 stands for "Text-to-Text Transfer Transformer." Developed by Google AI, T5 is also a large language

model (LLM) based on a pre-trained transformer architecture. It employs a self-supervised learning objective known as masked language modeling (MLM). Given a context or text passage along with keywords, the model is tasked with generating questions.
For question generation we applied the RQUGE metric [142] and the results are shown in a box plot in Figure 5.22 when single QG is applied and Figure 5.23 when Multiple QG is used.

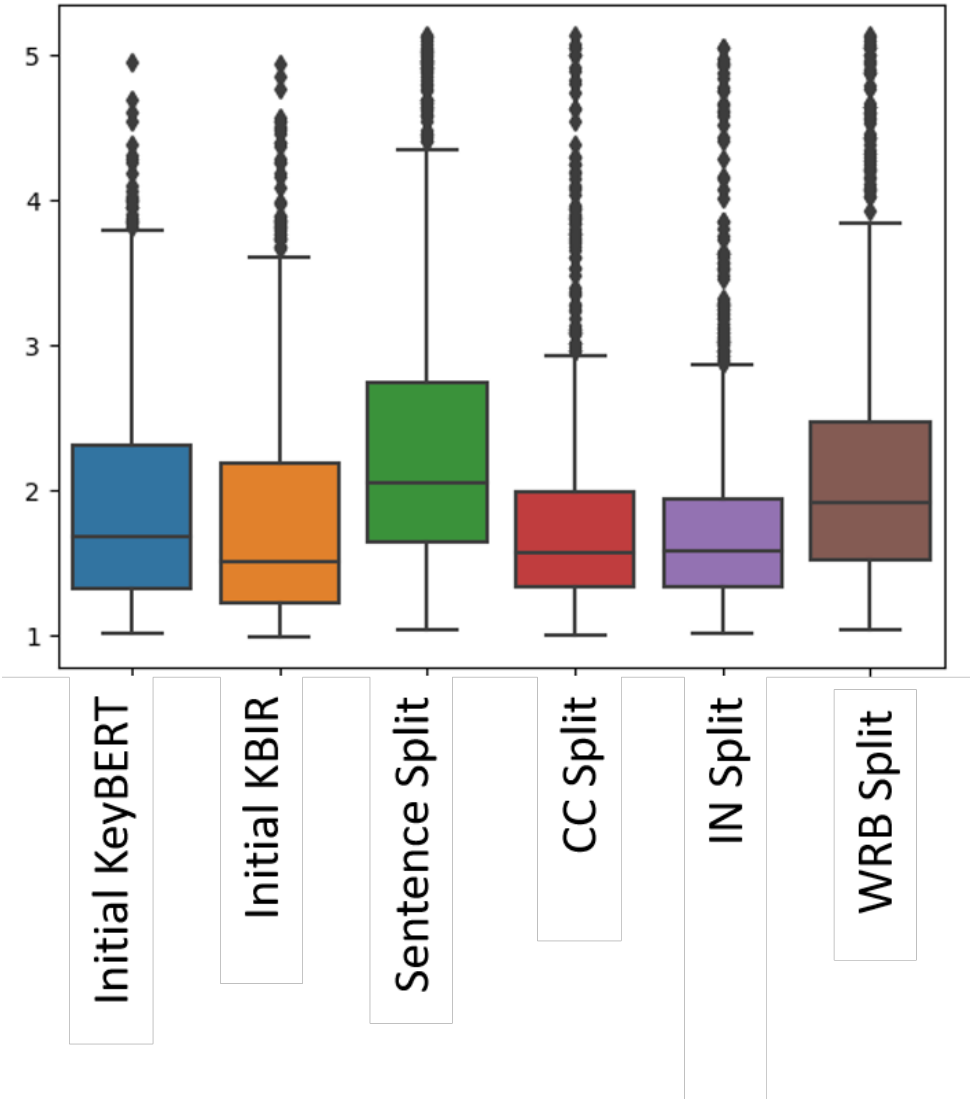


Figure 5.22: Box plot of RQUGE for Single Question Generation (SQG).

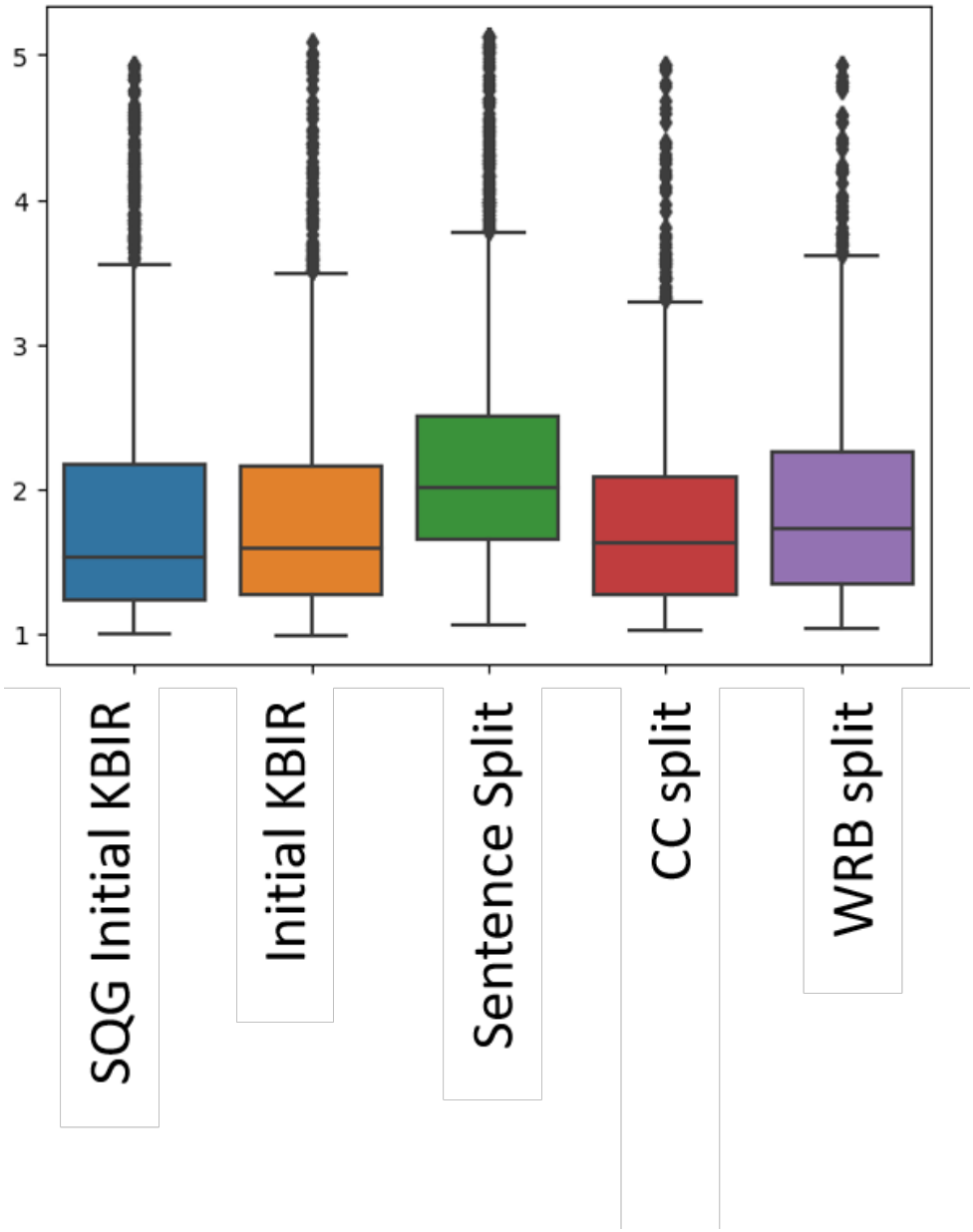


Figure 5.23: Box plot of RQUGE for Multiple Question Generation (MQG)

In our initial phase of question generation, our analysis showed a predominant use of the interrogative word "WHAT" in the questions generated. We counted interrogative words only if they appeared at the beginning of a sentence. We tested several configurations, and the following table outlines the variations observed concerning the types of interrogative questions generated. When utilizing T5 for multiple question generation through beam search, we noticed that the generated questions were strikingly similar, regardless of the temperature settings. As a result, we adopted a different approach for multiple question generation: we used sampling, set the most likely k sequences to 20, and then selected the top n, where n=8.

Regarding the quality of the questions, Table 5.4 displays the distribution of interrogative words for SQG, while Table 5.5 presents the same distribution for MQG.

Table 5.4: Question Diversity with SQG

	Keybert	**KBIR/INSPEC**				
		-	Sentence Split	CC split	IN Split	WRB Split
how	2	0	**219**	17	89	25
what	872	1890	2989	2758	**4023**	3974
when	1	0	28	11	**47**	16
where	3	7	**75**	1	20	3
whether	0	0	0	0	0	0
which	0	0	0	0	0	0
who	17	50	57	20	**65**	38
whom	0	0	0	0	0	0
whose	0	0	0	0	0	0
why	0	0	**47**	2	5	8

Table 5.5: Question Diversity with MQG

	KBIR/INSPEC				
	SQG	MQG	sentence split	CC split	WRB split
how	0	388	2492	**4587**	3492
what	1890	11928	19752	**79945**	65042
when	0	167	607	**1405**	1093
where	7	195	296	**1714**	1600
whether	0	2	5	**20**	16
which	0	512	464	**3701**	3076
who	50	822	833	**5942**	4967
whom	0	0	0	0	0
whose	0	6	13	**109**	87
why	0	90	578	**1155**	917

Question Answering

After obtaining the keywords and questions, the next step in our pipeline involves generating the corresponding answers. To achieve this, we employ a neural network technique using a pre-trained model named SciBERT [286]. SciBERT is a variant of the BERT language model that has been pre-trained on an extensive corpus of scientific literature, encompassing content from the biomedical domain. It was favored over other models such as BERT and BioBERT [289] for our purposes. The specific iteration of SciBERT we utilized is SciBERT-SQuAD-QuAC [295]. This model represents SciBERT fine-tuned for Question Answering using a combination of the SQuAD2.0 [296] and QuAC [297] datasets.
The BERTScore can be used to have some insight in this part. BERTScore is a metric for evaluating the quality of text generation, particularly machine translation. It works by first obtaining BERT representations of each word in the candidate and reference texts by feeding them through a BERT model separately. Then, it computes an alignment between the candidate and reference words by computing pairwise cosine similarity. This alignment is then aggregated into precision and recall scores, which are then aggregated into a (modified) F1-score weighted using inverse-document-frequency values. BERTScore has been shown to align better to human judgements in evaluating translation compared to existing metrics [136]. A score of 1 indicates a perfect match, meaning the generated text is identical to the reference

in terms of the contextual embeddings (as represented by BERT) while a score of 0 indicates no similarity between the generated text and the reference. Figure 5.24 and Figure 5.25 show the box plot for SQG and MQG strategies with differing data augmentation techniques.

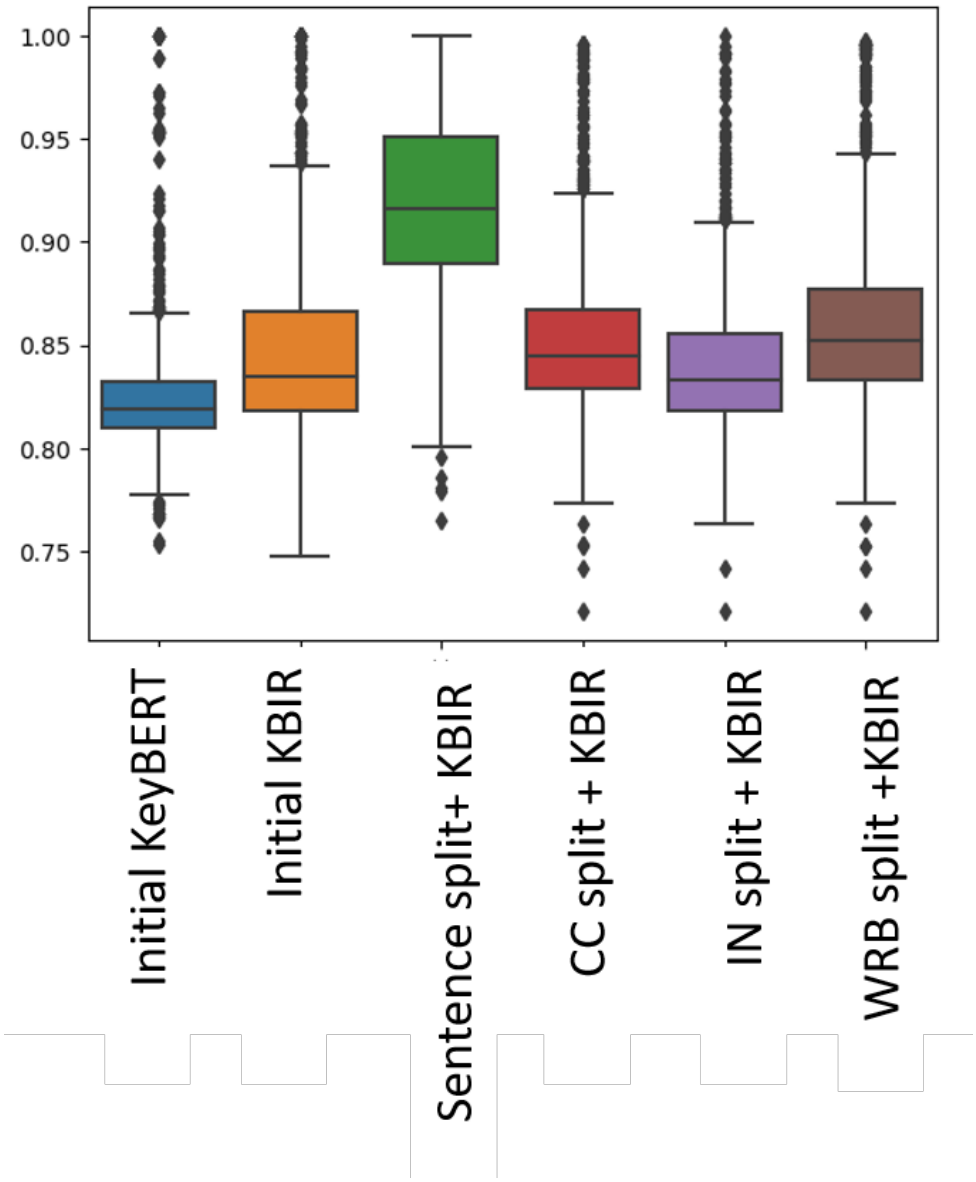


Figure 5.24: Box plot of BERTscore for Single Question Generation (SQG).

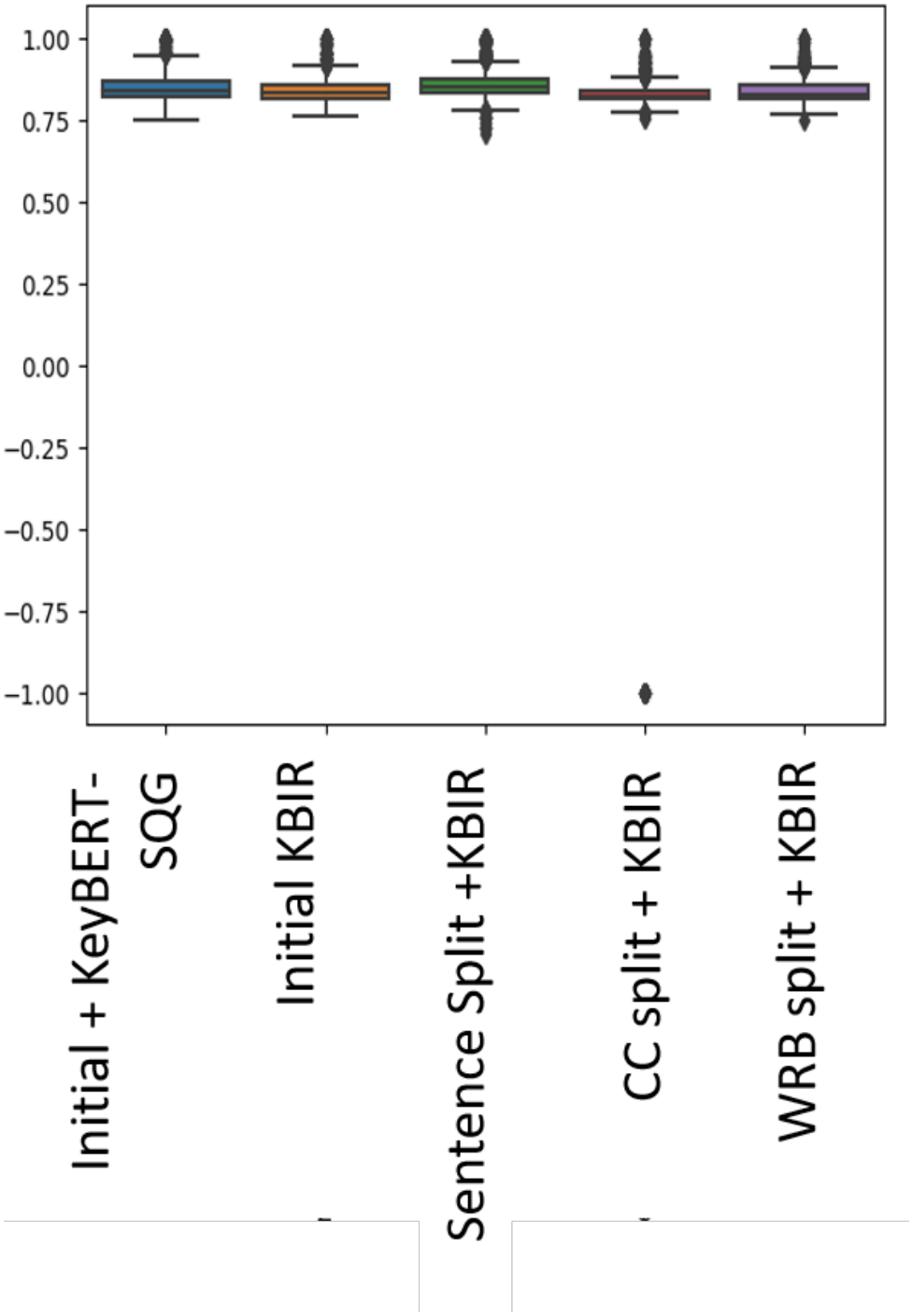
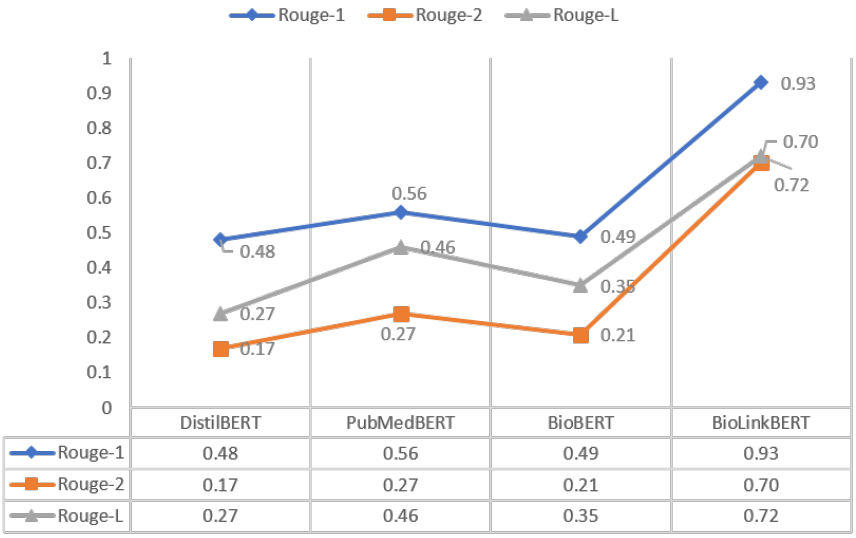


Figure 5.25: Box plot of BERTscore for Multiple Question Generation (MQG).

Chatbot Training

Figure 5.26 shows ROUGE scores for different models.



	DistilBERT	PubMedBERT	BioBERT	BioLinkBERT
Rouge-1	0.48	0.56	0.49	0.93
Rouge-2	0.17	0.27	0.21	0.70
Rouge-L	0.27	0.46	0.35	0.72

Figure 5.26: Rouge Metric.

In QA, Rouge-L can be used to measure the longest common subsequence between the generated answer and the reference (correct) answer. The longer the common subsequence, the better the answer is.
ROUGE-1 evaluates the overlap of unigrams, which are single words, between the generated text and the reference (correct) text. In other words, it measures how many individual words in the generated text are also found in the reference text.
ROUGE-2 assesses the overlap of bigrams, which are pairs of consecutive words, between the generated text and the reference text. It focuses on whether pairs of words in the generated text appear consecutively in the reference text.
The generated augmented dataset in the SQUAD-LIKE format was constituted by setting a threshold on the RQUGE score and F1-score as well. RQUGE was set to 1.5 and F1-score greater than 0. All triple sets of contexts, questions, and answers with an RQUGE and F1-score were added to the dataset used to fine-tune several pre-trained models, as shown in Figures 5.27, 5.28, 5.29 and finally 5.30

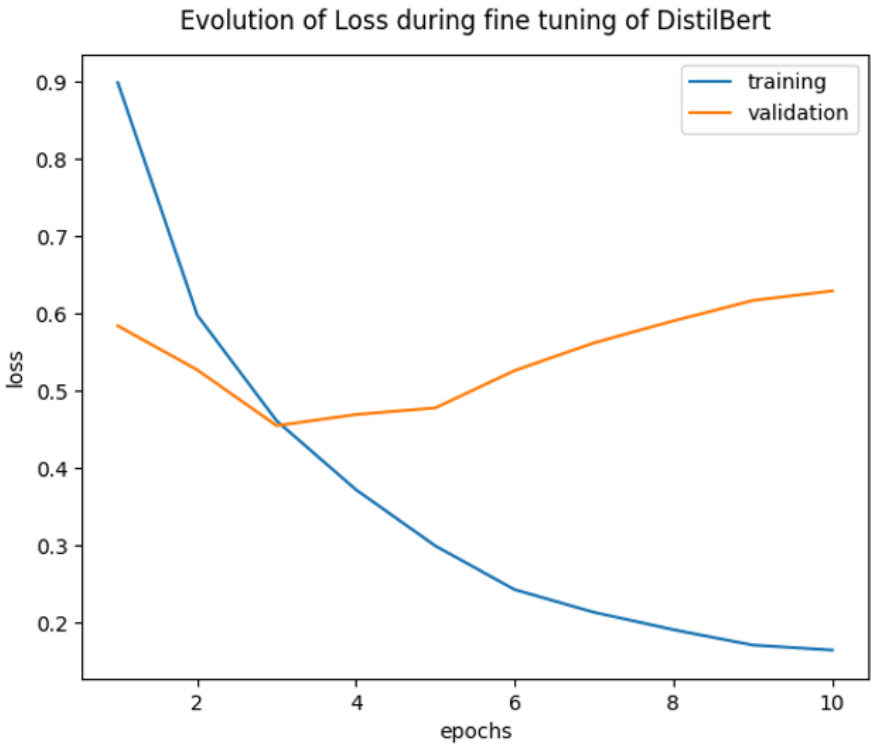


Figure 5.27: Loss Curve of Fine-tuning of DistilBERT.

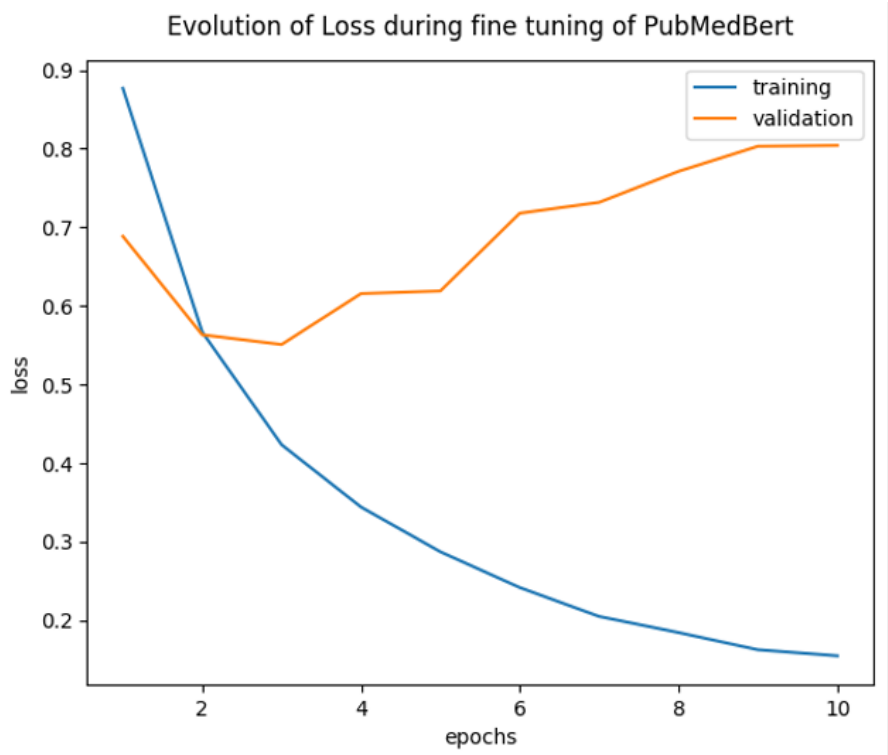


Figure 5.28: Loss Curve of Fine-tuning of PubMedBERT.

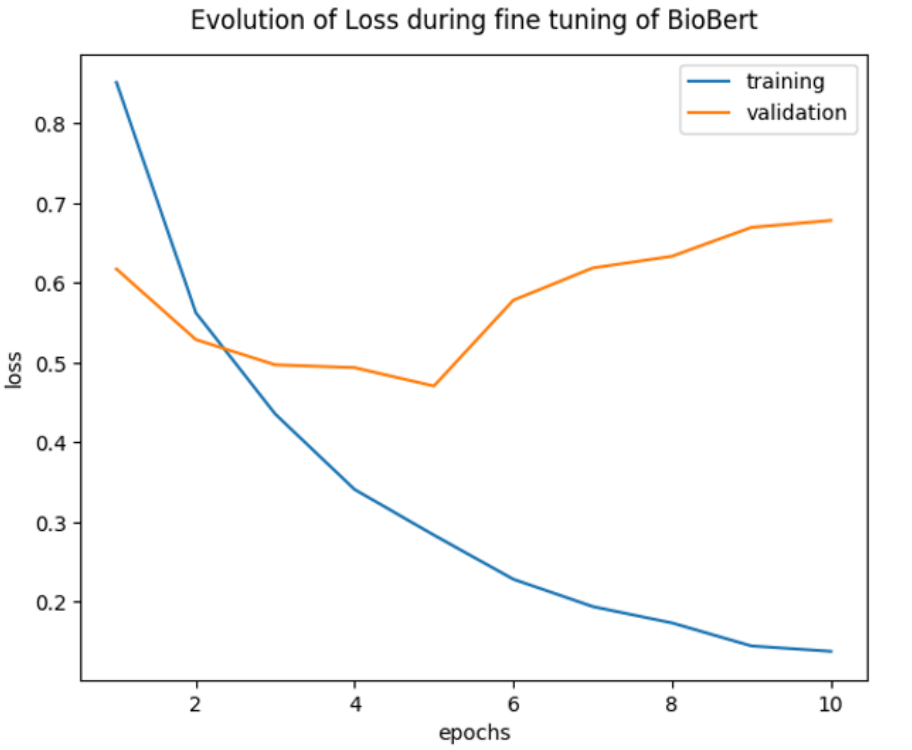


Figure 5.29: Loss Curve of Fine-tuning of BioBERT.

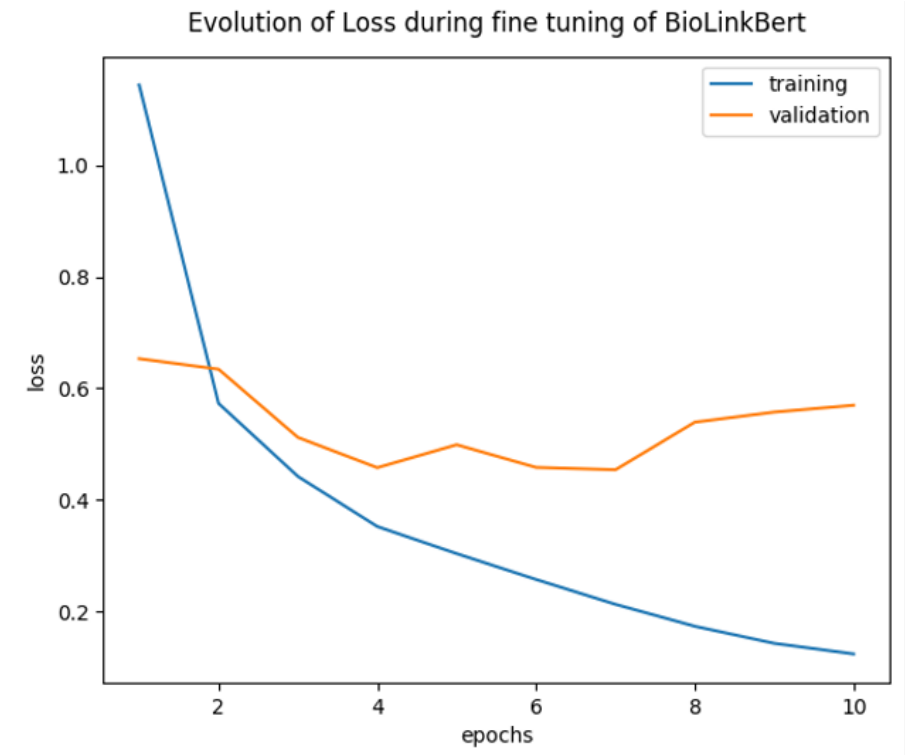


Figure 5.30: Loss Curve of Fine-tuning of BioLinkBERT.

5.5.4 Discussion

Our experiments on keyword extraction using the KBIR/INSPEC method, as shown in Tables 5.2 and 5.3, highlight how well our approach works in producing more keywords. It is important to mention that we chose not to use the KeyBERT model because it required a lot of computational power. By using different techniques to enrich our original dataset, we managed to greatly increase the number of keywords. For instance, when generating a single question, the keyword count rose from an initial 895 to 4066. For multiple question generation, the count jumped from 1948 all the way to a notable 111666 keywords. The introduction of Multiple Question Generation (MQG) in the KBIR model significantly boosts keyword generation, as evidenced from the Initial dataset comparison between SQG and MQG from Table 5.3. Dataset pre-processing strategies, especially "CC Split", lead to a substantial increase in keyword generation when combined with the KBIR model, as the "CC Split" produced the highest keyword count among the results. However, the choice of the dataset to use to train the chatbot is not solely dependant on number of keywords.

Table 5.4 and Table 5.5 show the distribution of different interrogative words used in questions generated through SQG and MQG. SQG and MQG strategies both lean heavily towards the usage of the interrogative "what". Furthermore, the "CC Split" strategy consistently produces higher counts across most interrogative types, indicating its tendency to generate a more diverse set of questions. Interrogatives like "whether", "which", "whom", and "whose" are used sparingly, suggesting that out dataset might have limited contexts prompting these specific question types. The prominence of "what" questions suggest that the dataset collected primarily cater to fact-based or descriptive queries, as opposed to more nuanced or detailed inquiries that might arise from "how", "why", or "which" questions. The indicated the need to have a more enriched and diverse dataset.

In comparison to existing research within the field, this work stands out as one of the pioneering efforts to utilize RQUGE as a metric for validating question generation within a chatbot pipeline. RQUGE serves as an evaluation metric designed specifically for assessing the quality of generated questions. What sets RQUGE apart is its ability to evaluate the quality of a candidate question independently, without the need for comparison to a reference question. It operates by considering the relevant context and answer span and employs a general question-answering module followed by a span scoring mechanism to determine an acceptability score.

From Figure 5.22 and Figure 5.23 box plots we can conclude that for both strategies, SQG and MQG, the Sentence Split and WRB Split methods exhibit a wider range of RQUGE scores, suggesting more variability in their question quality. The medians for these strategies are also generally higher, indicating that, on average, they produce better or more relevant questions. However, it is important to note the presence of outliers in all strategies, which could indicate occasional questions of either very high or very low quality. The RQUGE analysis

here does not give a clear idea of which dataset augmentation strategy will ensure our chatbot can be better train and hence perform better.
We now move to the analysis based on BERTscore from Figure 5.24 and Figure 5.25. For SQG sentence split strategy displays the most consistently high scores, ranging from about 0.90 to nearly 1.00 while for MQG all the strategies seem to perform at a very high level, with BERTscore close to 1. This suggests that questions generated through these strategies are likely highly accurate and contextually appropriate. However, since the scores are close it does not give us enough evidence as too which dataset is better suited for training the chatbot. However,we can see that the set without data augmentation is consistently lagging behind and in this case the CC split strategy has most adverse performance.
Figure 5.26 shows scores for four different BERT-based models: DistilBERT, PubMedBERT, BioBERT, and BioLinkBERT. We can make the following observations for Rouge-1:

1. BioLinkBERT significantly outperforms the other models with a score of 0.93.
2. PubMedBERT comes second with a score of 0.56.
3. DistilBERT and BioBERT are close, with scores of 0.48 and 0.49 respectively.

For Rouge-2:

1. Again, BioLinkBERT has the highest score of 0.70, indicating a greater overlap of bigrams with the reference.
2. PubMedBERT follows with 0.27.
3. BioBERT scores 0.21 and DistilBERT is at the bottom with 0.17.

For Rouge-L:

1. BioLinkBERT leads with a score of 0.72.
2. PubMedBERT scores 0.46.
3. BioBERT is close behind with 0.35, and DistilBERT has the lowest score of 0.27.

BioLinkBERT consistently achieves the highest scores across all three ROUGE metrics, indicating that its generated answers having the highest overlap in relation to keywords in terms of unigrams, bigrams, and the longest common subsequence. This suggests that BioLinkBERT might be the most effective model among the four, for generating answers, in contexts where the ROUGE metric is relevant.
Furthermore, the training loss and validation loss curves we notice that, for all models, training loss decreases consistently, which is expected as the models adjust their weights to minimize the error on the training dataset. However, validation loss for all models initially drops, suggesting that the learning is generalizing well. However, after a certain number of epochs, the validation loss starts to rise for most models, indicating overfitting. This is where the models are becoming too specialized to the training data and are losing their generalization capability on unseen data.
Contrasting with the approach in [265], where question answering is conversation-based, our solution is tailored for a closed domain, with question answering based on a dataset comprising context, questions, and answers. Additionally, [265] employed the F1-score metric and introduced human intervention for performance comparison. Moreover, they had access to

gold-standard answers for the questions. Comparing our work with that of Chan *et al.* [269], a similarity lies in the utilization of pre-trained BERT language models. However, the distinction lies in our extensive experimentation and testing of various pre-trained BERT model variations. While they employed beam search, we explored this approach and ultimately decided against its use for question generation due to the limited diversity in the generated questions.
The fine-tuning of various BERT-based models for the chatbot reveals that BioLinkBERT exhibits significantly better performance, as depicted in Figure 5.26. The results presented in this chapter represent the first iteration of our overall testing process, which is still ongoing. There are additional data augmentation techniques that can be explored, as well as a range of models that were not tested due to significant processing power limitations. Nevertheless, for the final step of the pipeline, which involves fine-tuning a pre-trained model with our dataset, BioLinkBERT appears to be the most suitable choice based on observed results. However, based on the overall results observed, it would indicate that the BioLinkBERT model is better suited for our task. However, concerning which final dataset needs to be used for fine-tuning, this is still open to further exploration and experimentation.

5.5.5 Limitations

In our research, a significant limitation arose from the lack of comprehensive contextual information pertaining to all facets of diabetes, which was necessary for constructing our dataset. Additionally, we were constrained by limited computational capacity. While we tried to address the data limitation challenge using data augmentation techniques, we were faced extended computational delays. Specifically, a single cycle of data augmentation necessitated a minimum of two weeks. Consequently, this restricted our ability to undertake more extensive experimentation, which, in our assessment, could have enhanced the efficacy of our processing pipeline.

5.6 Conclusion and Future Works

This initial research represents a modest yet significant step towards establishing an AI-driven ecosystem to combat diabetes. Given the right dataset and sufficient computational resources, this system has the potential to significantly alleviate the burden on healthcare professionals by addressing routine, repetitive inquiries, thereby allowing them more time to focus on cases that necessitate human intervention. Recommended areas for future exploration include:

1. Enhancing the dataset by incorporating authentic questions posed by patients and the corresponding responses provided by healthcare professionals. By tailoring the implementation to specific geographical locations, the system can capture and integrate local linguistic nuances, ensuring a richer and more contextually relevant dataset.

2. Incorporating discussions between patients and healthcare professionals into the dataset is a highly sensitive task, fraught with ethical considerations. It is most important to research and apply robust anonymization techniques to ensure the protection of individual privacy and to uphold the highest ethical standards in the process.
3. Incorporating a Neural Information Retrieval (NIR) component can be instrumental in continually enhancing the contextual dataset. NIR, a specialized branch of information retrieval as described in [298] and [299], leverages neural network models and methodologies to refine the retrieval of relevent information from vast document collections or datasets. By integrating this component, it becomes possible to achieve more accurate and contextually rich results, driving the efficacy of the overall system.
4. Integrating an expert sub-system can significantly enhance the proactive capabilities of the chatbot. This automated approach would enable the chatbot to analyze a patient's queries over time, thereby discerning distinctive "profiles" or patterns. Based on these recognized patterns, the chatbot could then proactively engage patients by posing specific targeted questions. The objective of this interaction is to either validate or challenge a formulated hypothesis, ensuring a more personalized user experience.
5. We recognize that in its early stages, the chatbot might face challenges in delivering satisfactory answers owing to the constraints of the initial dataset. To address this shortcoming, we recommend the incorporation of a question storage mechanism. Additionally, the adoption of active learning strategies [300] could be advantageous. By utilizing these techniques, unanswered queries can be annotated by a panel of healthcare experts. Such a methodology not only promises to improve the chatbot's efficiency but also significantly expand its repository of knowledge.

This research presents a new chatbot training pipeline that takes advantage of pre-trained Transformer-based models. We have adopted RQUGE, a recent metric for evaluating NLP systems, which has given us important insights into our system's performance. Additionally, using the KBIR/INSPEC model for keyword extraction has shown good results compared to the highly effective KeyBERT but computationally intensive pre-trained model.

6

Conclusion and Perspectives

Summary

6.1 Introduction

Artificial Intelligence (AI) stands as a pivotal digital health technology poised to revolutionize the landscape of diabetes care, as highlighted by Klonoff *et al.* [301]. In this thesis, we have embarked on a journey to enhance a traditional healthcare ecosystem, known as the Learning Nest (LN) [46], in the context of diabetes prevention, education, and self-management. Our approach has centered on the application of AI to both image and text data. The research endeavors undertaken within this thesis have spanned various domains, including the management of Diabetic Foot Ulcers in chapter 3, an exploration of the confidentiality issues surrounding health-related data in 4 and finally the development of an AI Chatbot for diabetes-related inquiries in chapter 5. Hereafter, a summary of the contributions is presented, and the overall findings are discussed, followed by the final section elaborating on potential future investigations and open challenges.

6.2 Summary of Contributions

6.2.1 DFU-SIAM

We conducted a comprehensive study in the field of Diabetic Foot Ulcer (DFU) classification, which had previously relied on the Kaggle Dataset and primarily focused on binary classifications such as normal vs. abnormal or infection vs. non-infection classes. We followed this up by studying the latest research which uses the meticulously labeled dataset acquired from the DFU 2021 challenge [195] which classifies DFU into four distinct classes. We thus propose a multi-class classification on DFU which outperforms all the previous research mentioned in chapter 3. Recognizing the significant class imbalance in the dataset, we applied geometric transformations to augment the DFU images. For this step experimentation were also conducted to ensure geometric conversion which positively impacted performance were kept. Our novel approach involved creating an ensemble of CNN-based pre-trained architectures, specifically EfficientNetV2S [171] and Vision Transformer (BEiT) [172], [178], deployed in a Siamese Neural Network (SNN) architecture. Unlike conventional research which applies contrastive loss functions when using SNN, we propose to use an innovative Large Margin Cotangent Loss (LMCoT) [173], which utilizes the cotangent function instead of the cosine function to

implement the margin function. To further enhance our model's performance, we implemented a K-Nearest Neighbors (KNN) classifier. We iteratively determined the optimal value of K for each epoch based on the F1-score, ensuring that our model achieved the best results when classifying unseen DFU images, surpassing previous research based on the DFU2021 challenge dataset. Finally, we emphasized the critical need for interpretability in machine learning models used in healthcare. While achieving accurate DFU classification marks significant progress, we also outlined avenues for future work, including employing this model for DFU follow-up and validating DFU management protocols.

6.2.2 DFU-Helper

Building upon our initial investigations, we embarked on a deeper exploration to fully leverage the findings from our earlier research. While our initial work represented a significant breakthrough in the classification task, we were driven by the desire to make a meaningful impact within our ecosystem. Hence, we decided to investigate the deployment of a system that would consolidate the insights of healthcare professionals who initiate treatment protocols upon detecting DFUs. This led us to propose the development of a longitudinal evaluation system for the disease. We once again leveraged the DFU2021 dataset, augmenting it with additional images and introducing a "Healthy" class. The additional images were obtained from Kaggle. We thus have a dataset with 5 classes. We used the similarity learning capabilities of SNN to create class anchors for all our classes and plotted the similarity of a new DFU image along five axes in a radar plot, using both cosine and Euclidean distances. This approach enabled us to track progress over time across these five axes at different time periods. With this method, the diagnosis of positive or negative progress is no longer solely subjective but is supported by this automated solution, with the medical practitioner remaining the primary decision maker. Our next steps involve achieving a balanced dataset, expanding our dataset comprehensively, and, most importantly, ensuring the implementation of explainability (XAI) for the model's outcomes.

6.2.3 Confidentiality of healthcare data

To train a model we need good reliable data which are not always available at the same geographical site. One of the recommendation of our work on DFU to curb dataset limitation was to explore data available from different healthcare entities working on diabetes management and foot care. However, there is an issue of sharing sensitive personal medical data of patient which poses ethical, privacy and confidentiality issue. We thus study Federated Learning (FL) [183], a machine learning technique that enables multiple devices to collaboratively train a model without sharing their raw data. This process involves each device training the model on its local data and then transmitting the updated model parameters to a central server. The central server aggregates these updates and sends back the improved model to the devices.

This iterative process continues until the model converges. In our experiments, we employ a siamese model and the DFU dataset [195]. We evaluate the performance of both centralized FL and Peer-to-Peer FL [250] by applying different heuristics for node selection. Our findings reveal that the Peer-to-Peer FL model achieves accuracy levels comparable to those of the centralized FL learning architecture. However, further exploration involving additional heuristics and more robust processing power is required to ensure comprehensive experimentation in this area.

6.2.4 AI Chatbot for Diabetes [AICHAD]

We introduce a novel chatbot pipeline based on a pre-trained variant of the Transformer model [15], specifically the BERT variant [44]. The underlying hypothesis of this work relies on the ability of transformers to grasp semantic meanings and contextual nuances in language. Our goal is to transform this proposed chatbot into an advanced conversational agent that can adeptly capture and adapt to the of language usage in diverse local contexts worldwide, leveraging datasets derived from these geographical spheres. The pipeline we propose is built upon a dataset encompassing context, questions, and answers, with a specific focus on diabetes-related information. To enhance the dataset's quality and performance, we employ several data augmentation techniques. We address the computational complexity associated with keyword extraction by replacing KeyBERT [280] with KBIR-INSPEC [281]. Furthermore, we explore question generation using the Text-to-Text Transfer Transformer (T5) model [285], encompassing both single and multiple question generation. Our evaluation of these questions centers on assessing their qualitative aspects, including the diversity of interrogative words used. We implement the most recent metric for question assessment, known as Reference-Free Question Generation Evaluation (RQUGE) [142]. RQUGE evaluates the quality of automatically generated questions without the need for reference questions, and it has demonstrated strong correlations with human judgments of question quality. One of its key principles is that a good question should be answerable and informative. We set a threshold for RQUGE, selecting questions with scores greater than 1.5 for inclusion in our final dataset. For answers generation we employ the pre-trained model SciBERT [286], fine-tuned on the SQuAD2.0 [296] and QuAC [297] datasets. Finally, we train our chatbot model by fine-tuning our dataset on various BERT-based models, including DistilBERT [290], PubmedBERT [288], BioBERT [289], and BioLinkBERT [291]. Among these models, BioLinkBERT consistently exhibits superior performance. It is important to note that there are still more models and experiments to be conducted in relation to the research done. Future phases of this project will involve advanced techniques such as neural information retrieval, active learning, and annotation transformers, marking a promising avenue in the field of AI chatbots to support prevention, education, and management of chronic diseases in healthcare. As a final note, we leveraged the creation of the diabetes dataset from media articles and conducted a survey applying natural language

techniques. The results are presented in Annex A.1.

6.3 Discussions and Limitations

While AI has been extensively discussed in the literature as a transformative tool in the healthcare industry, this thesis afforded the opportunity to explore several of the latest techniques based neural network architecture approaches, including pre-trained CNNs, EfficientNet [152], VIT [115], SNN [124], and BERT [44]. However, the practical implementation of these AI solutions brought certain challenges to the forefront. Chief among these challenges is the availability of suitable datasets, closely followed by the accessibility of sufficient processing power. A critical problem in deep learning is that systems learn inappropriate biases, resulting in their inability to perform well on minority groups [302]. Concerns regarding potential biases and limited generalization capabilities across various demographic factors, such as gender, age distributions, races, ethnicities, hospitals, and variations in data acquisition equipment and protocols, have been extensively highlighted in the literature [303]. In the context of our research on DFU, we encountered imbalanced data distribution among certain classes, particularly in the "both" and "ischaemia" classes. This imbalance poses a significant challenge and necessitates further attention. Future efforts should focus on expanding data collection and exploring advanced data augmentation techniques to rectify this issue. Additionally, the incorporation GANS warrants further investigation in the DFU research to promote a more balanced dataset, which can ultimately enhance the effectiveness of AI-based solutions in this domain. By further researching the federated learning architectures we can also encourage collaboration between distributed entities to collaborate for data collection by ensuring safeguard of confidentiality. The availability of appropriate dataset was also an issue in the research that involved the Mauritian Dish recommendation and conversational agent.
In the context of the conversational agent, the absence of sufficient fine-tuning data in a specific domain can significantly affect the chatbot's capacity to generate accurate responses. For instance, if the chatbot lacks exposure to contextual information related to gestational diabetes, it will consistently provide incorrect answers to questions related to this specific topic. Addressing the challenge of dataset availability through data augmentation techniques also brought to the forefront the issue of processing power. As an illustration, the application of a part-of-speech tagging strategy for data augmentation on the initial context dataset for Question Generation (QG) required an extensive amount of computational resources. The execution of this process took up to 30 days to complete if no memory issue occurred. Testing alternative combinations of pre-trained architectures for the DFU research with SNN proved to be challenging due to the limitation of GPU memory. The constraints on available GPU memory prevented the exploration of various model configurations and architectures, which can be

crucial for optimizing the performance of AI models. This was also a limiting factor for testing the required architecture and the number of clients in P2P federated learning environment. Within the scope of the various research endeavors pursued, we acknowledge the need to do extensive research into the field of Explainable AI (XAI), which is considered a critical component for user acceptance in the medical domain [189]. XAI tries to ensure that AI algorithms (and the resulting decisions) can be understood by humans [304]. It plays an indispensable role in providing transparency and interpretability to AI models, especially in healthcare applications where decisions have a direct impact on patient well-being. Although it would have been beneficial to consider the XAI aspect of the applied models across our studies, the primary focus of this thesis was on the predictive performances.

6.4 Conclusion and Perspectives

This thesis investigated the application of AI techniques within a traditional ecosystem aimed at preventing, educating, and managing T2D through the utilization of image and textual data. Nevertheless, several research challenges have surfaced during the course of this study, and they are outlined as follows for future research considerations:

1. Further investigation into Federated Learning algorithms, both centralized and distributed, for facilitating data sharing during the training of deep learning models and ensuring confidentiality.
2. The development of a sophisticated tool for longitudinal DFU follow-up is a time-consuming process that necessitates an iterative development process, incorporating feedback from healthcare practitioners. The next step involves collaborating with experts in the field of public health to design an evaluation protocol that will be implemented with the collaboration healthcare practitioners. The ultimate goal is to deploy this tool as a self-care application on mobile devices for both patients and their family members. It is crucial to emphasize that a major focus in this future work should be placed on ensuring explainability of the system.
3. Diabetic Retinopathy (DR) is a diabetes-related complication that causes visual impairment. Screening for diabetic retinopathy (DR) is recommended to detect sight-threatening complications prior to visual loss [305]. Complications can be avoided by early detection and various deep learning models, trained on numerous annotated retinal images [306]. Extend the application of the DFU classification model and DFU longitudinal evaluation to the fields of Diabetic Retinopathy (DR) and Nephropathy.
4. Investigate the implementation of deep learning techniques, such as YOLO and others, for calorie estimation, with a specific focus on Mauritian food dishes using mobile phones or explore other avenues as the smart plate project [307]. As a starting point implementing

this approach for calorie calculation in fruits using the recently available DeepFruit dataset [308] can be considered.

5. Revise the extractive conversational agent to include real-time dialogue monitoring, incorporating the entire current dialogue as context to deduce user intent and generate responses. Further, extend the research to enhance the answering component of the chatbot and evaluate its performance using a broader range of pre-trained models. It is imperative to prioritize the ethics and privacy of information, especially in this context where patients will be sharing personal data with the system. These data will be utilized to enhance the system's answer generation capabilities. The WHO ethics [66] recommendations should be the guiding principle.

This research stands as a pioneering attempt in the creation of an AI-powered ecosystem designed to address the multifaceted challenges of prevention, education, and management of diabetes. It is of utmost importance to clarify that this work does not advocate for clinical diagnosis or the replacement of healthcare professionals; rather, its central aim is to equip the healthcare sector with AI tools to enhance patient outcomes. It is both acknowledged and demonstrated that while the availability and utilization of cutting-edge technologies offer substantial benefits to individuals living with diabetes, they do not eradicate the disease itself [309]. Nevertheless, the advantages in terms of cost-efficiency, personalized self-care, reduced burdens on healthcare experts, and the associated socioeconomic factors underscore the need for continued research and commitment to leveraging AI in the fight against diabetes. The WHO has unequivocally recognized AI's potential to address various healthcare sector challenges [20]. Confronting the herculean task of combating diabetes requires transformative changes beyond the scope of any single policy or intervention [310]. It necessitates a multidisciplinary and innovative approach, and we are proud to have contributed "d'avoir apporter notre pierre à l'édifice" towards this critical objective.

A

Annexes

Summary

During the work carried out in this doctoral thesis, I was led to explore other avenues that allowed me to conduct two studies:

1. An NLP-based study to compare the textual content of online news media related with diabetes in France and the Indian Ocean to highlight the differences in dealing with diabetes between these two regions.
2. Image data augmentation for the implementation of an effective classification model for the detection of Mauritian dishes.

The research conducted around these two themes has led to two scientific publications, which are presented in the Section A.1 and Section A.2

A.1 NLP-Media Article

Analyzing Online Media Articles on Diabetes using Natural Language Processing: A Comparative Study of Indian Ocean Region and France

Mohammud Shaad Ally Toofanee[1,2], Nabeelah Zainab Ally Pooloo[2], Sabeena Dowlut[2], Karim Tamine[1], and Damien Sauveron[1]

[1] XLIM, UMR CNRS 7252, University of Limoges, 123,Avenue Albert Thomas, 87060 Limoges, France
[2] Université des Mascareignes, Concorde Avenue Roches Brunes Rose Hill, Mauritius

1 Abstract

Background: Diabetes is a global health concern affecting millions of people worldwide. However, knowledge, attitudes, and practices related to this disease vary widely across different regions. This article aims to investigate media-influenced perceptions about diabetes in France and the Indian Ocean countries using natural language processing (NLP) techniques applied to online news articles. Findings aims to provide expert in Health Literacy (HL) and health promotion to develop better communication strategies. **Method**: Constitute a datatset of Online news articles on Diabetes and apply NLP like Word2Vec for word integration, LDA for topic identification, and transformer-based classification models (e.g., BERT and its variants) for sentiment analysis. processing (NLP). **Results**: Sentiment analysis revealed more negative discussions about diabetes in the Indian Ocean region (48%) compared to France (32%), with neutral articles dominating in France (42%). In terms of topic Identification there were some topic which appeared for France which were not present for indian ocean region. **Discussions**: The findings of this study indicate that perceptions and discussions about diabetes differ between two regions, which have implications for public health interventions and communication strategies. However, the study is limited by the initial amount of information captured for analysis.

Keywords: Artificial Intelligence, Natural Language Processing, Mass Media, Diabetes, LDA,Transformers, BERT, Sentiment Analysis, Word Associations

Analyzing Online Media Articles on Diabetes using Natural Language Processing: A Comparative Study of Indian Ocean Region and France

Mohammud Shaad Ally Toofanee[1,2], Nabeelah Zainab Ally Pooloo[2], Sabeena Dowlut[2], Karim Tamine[1], and Damien Sauveron[1]

[1] XLIM, UMR CNRS 7252, University of Limoges, 123,Avenue Albert Thomas, 87060 Limoges, France
[2] Université des Mascareignes, Concorde Avenue Roches Brunes Rose Hill, Mauritius

1 Abstract

Background: Diabetes is a global health concern affecting millions of people worldwide. However, knowledge, attitudes, and practices related to this disease vary widely across different regions. This article aims to investigate media-influenced perceptions about diabetes in France and the Indian Ocean countries using natural language processing (NLP) techniques applied to online news articles. Findings aims to provide expert in Health Literacy (HL) and health promotion to develop better communication strategies. **Method**: Constitute a datatset of Online news articles on Diabetes and apply NLP like Word2Vec for word integration, LDA for topic identification, and transformer-based classification models (e.g., BERT and its variants) for sentiment analysis. processing (NLP). **Results**: Sentiment analysis revealed more negative discussions about diabetes in the Indian Ocean region (48%) compared to France (32%), with neutral articles dominating in France (42%). In terms of topic Identification there were some topic which appeared for France which were not present for indian ocean region. **Discussions**: The findings of this study indicate that perceptions and discussions about diabetes differ between two regions, which have implications for public health interventions and communication strategies. However, the study is limited by the initial amount of information captured for analysis.

Keywords: Artificial Intelligence, Natural Language Processing, Mass Media, Diabetes, LDA,Transformers, BERT, Sentiment Analysis, Word Associations

2 Introduction

The internet has become the predominant source of information for individuals when encountering various problems or health conditions. A recent study concluded that 74.4% of individuals in the United States initially sought health-related information online, while only 13.3% consulted a healthcare professional as their first step[1]. However, online platforms often contain poorly reported information, including misinformation, cherry-picked data, exaggerated claims, and other misleading content, posing a significant risk to public health [2]. This issue has been widely observed across different regions of the world, particularly in the context of the COVID-19 pandemic. Additionally, previous research has highlighted the role of mass media communication in shaping public opinions[3]

Type 2 Diabetes (T2D) is a significant health concern in Mauritius, with the International Diabetes Federation (IDF) predicting a prevalence rate of 26.6% by the year 2045 [4]. In this study, we propose utilizing artificial intelligence methods, specifically natural language processing (NLP), to analyze online media articles discussing T2D in countries within the region and compare them with a developing country in Europe, namely France. Our objective is to provide health professionals with some additional inputs to adapt communication message and strategies for T2D prevention and education in these regions.

NLP enables the exploitation and manipulation of a vast amount of data to gain insight that would otherwise not be humanly possible. However, NLP tends to confirm clinical hypotheses rather than develop entirely new information [5]. while several researches has been carried out on information exchanged on social media platforms, mainly twitter, this is not possible for the local context. In the local context talking about a pathology remains taboo hence the important to analyse media topics and sentiments since it is one of the main source of information on health related issues and the impact in has on public opinion The dataset constituted from this research is also intended to be used as input for a project on AI powered chabot for diabetes prevention and management.

Natural Language Processing (NLP) as a powerful tool for extracting and analyzing large amounts of data to gain insights that may not be easily attainable through human efforts. While many studies have utilized social media platforms, such as Twitter, for health-related information analysis, this approach may not be feasible in local contexts where discussing certain health issues is considered taboo. Thus, it is essential to analyze media topics and sentiments as they serve as a valuable source of information regarding health issues and their influence on public opinion.

Based on both psychology and sociology, the framing effect theory explains the ability of news media to affect people's attitudes and behaviors through making slight changes[6] The findings of this comparative study can provide valuable insights for designing effective communication strategies and interventions to address the complex social and psychological dynamics associated with diabetes.

This work focuses on three key text mining tasks: sentiment analysis, automated topic extraction (Topic Modeling), and semantic correlation of words with their context related to diabetes in online news media. The dataset generated from this study is intended to be used as part of the training for an AI-powered chatbot project aimed at diabetes prevention and management.

3 Related Works

Deep learning has being widely utilized in the medical field, particularly in the analysis of medical digital images. However, there has been a recent trend towards exploring the potential of textual data as well. In a recent study, Boissonnet et al. (2002) proposed a model for evaluating the quality of health-related articles and providing explanations for the classification they make, utilizing the BERT (Bidirectional Encoder Representations from Transformers) approach [2]. This work contributes to the growing body of research on leveraging NLP techniques for analyzing textual data in the context of health-related articles.

In their recent publication, UnKyo et al. (2002) conducted a study on sentiment analysis of telemedicine-related newspaper articles during the COVID-19 pandemic in Korea, investigating the association between the pandemic and changes in the media's perception of telemedicine. They employed Latent Dirichlet allocation (LDA) analysis, topic extraction, and topic trend analysis to analyze the data and draw conclusions [6]. This research contributes to the understanding of how the perception of telemedicine has evolved during the pandemic.Furthermore,in a separate study, Wang et al. (2022) aimed to demonstrate the applicability of natural language processing (NLP) techniques in the Chinese language medical environment. They successfully utilized NLP techniques to rapidly identify vulnerability factors in the management of Type 2 diabetes (T2DM) [7]. Their research highlights the potential of NLP in uncovering insights in the Chinese language medical domain.

In the study conducted by Oyebode et al. (2019), they applied natural language processing (NLP) techniques to social media data collected from Nigerian platforms to detect factors responsible for diabetes prevalence [8]. Using the Binarized Naïve Bayes (BNB) algorithm, their solution revealed significant factors such as weight, diet, pregnancy, age, and sleep that contribute to diabetes prevalence. These findings provide valuable insights for actors in the health sector to guide interventions in diabetes education and prevention efforts.Building upon their previous work, the same authors developed a medical named entity recognition framework called MediNER, which effectively identifies named entities related to diabetes management and classifies them into categories such as Food, Medication, Therapeutic Procedure, and Supplement [9]. This research showcases the potential of NLP techniques in improving diabetes management through precise identification of relevant entities.

The mining of data exchanged on social media platforms regarding COVID-19 and vaccination has also gained significant attention from the research commu-

nity, with studies conducted by Praveen et al. (2022), Canaparo et al. (2023), and Zulfiker et al. (2022)[10,11,12].

Additionally, several researchers have applied NLP techniques for sentiment analysis related to diabetes and social networks, as demonstrated in the works of Gabarron et al. (2019), Salas et al. (2017), De et al. (2012), and Liu et al. (2020) [13,14,15,16]. These studies collectively highlight the increasing interest in utilizing NLP techniques to gain insights from social media data in the context of diabetes and related health issues.

Foley et al.(2020) researched how diabetes pandemic was portrayed in the United kingdom news bewteen 1993-2013[17] and furthermore Syafhan et al. investigated on media reporting of antidiabetic medicines in newspapers published in the United Kindgom and United states[18]. Such endeavors hold immense potential for NLP experts to efficiently analyze vast amounts of information and generate numerous insights, thus making it a crucial area of research.

4 Materials and Methods

The figure 1 presents an overview of the tasks that need to be completed to achieve the objectives initially set with the first step being data collection. Each steps mentioned are explained in details in the following sections.

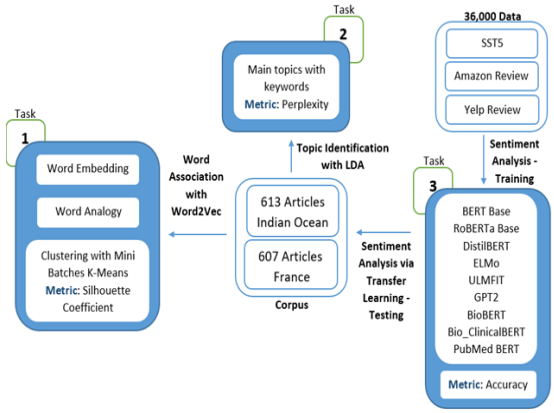


Fig. 1: Overview of Task to be completed

4.1 Dataset and Pre-processing

The first step involved collecting textual data from online newspaper publications in two regions, the Indian Ocean and France, in order to analyze specificities related to Type 2 Diabetes. We used web scraping techniques, such as Beautiful

Soup, ParseHub[19], and Octoparse[20], to extract electronic text data from structured web pages of news outlets and health magazines focused on diabetes. The collected data was consolidated and filtered to include only articles that were less than five years old to ensure relevance to current lifestyle practices and diabetes conditions. Pre-processing techniques were then applied to prepare the textual data for analysis.

1. Removing duplicate cells, empty values, punctuation marks, URLs,
2. Tokenisation, which is the process of dividing the text into smaller units,
3. Stemming, is a natural language processing technique that lowers inflection in words to their root forms,
4. Lemmatisation, which is the process of ensuring that etymological words do not lose their meaning.

4.2 Vectorisation, word association and word analogy

After data collection and pre-processing, Word2Vect method was applied for vectorisation. Word2vec is a technique of natural language processing [21] that is used to produce word embeddings. Word2vec takes as its input a large corpus of text and produces a vector representation, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space [21]].

We utilized the Word2Vec algorithm to implement word embedding, and performed word analogy analysis to gain insights into word associations. Popular libraries such as Keras and Gensim were employed for these tasks, with a focus on Gensim for its ability to visualize word embeddings and identify similar words. Additionally, the performance of the word embedding was evaluated through vector arithmetic-based word analogy examinations. The findings of this study shed light on the semantic relationships between words in the context of diabetes, providing valuable insights for further research in the field.

4.3 Latent Dirichlet allocation (LDA) analysis , Topic Identification and Modeling

Latent Dirichlet Allocation (LDA) approach was used to analyze two corpora related to diabetes in order to identify key topics addressed in the texts[22]. LDA is a generative statistical model commonly used for topic identification in text data. Topics are defined as groups of representative words that aid in identifying the subject matter of the text. LDA represents documents as a mixture of latent topics, where each topic is characterized by a distribution over words[23]. Topic identification is crucial for clustering documents, extracting information from unstructured text, and selecting features. In this study, we aimed to uncover

different topics, each represented by a combination of keywords, in the diabetes corpora to gain insights into the main themes discussed in the texts.

4.4 Sentiment analysis

Finally, we did sentiment analysis of the text collected from online media sources related to diabetes. We classified the sentiments into three possible forms: neutral, positive and negative. This was realized by setting up a classifier on textual data with three classes.

Due to the lack of a labelled dataset for diabetes, another alternative was used to train the NLP models for sentiment analysis. Three different and famously used datasets are retrieved from the internet, namely: SST5 [24], Amazon Review [25], and Yelp Review [26]. SST5 is the Stanford Sentiment Treebank dataset consisting of 5 classes of sentiment on movie reviews; it is well-regarded as a crucial dataset and used as a primary benchmark dataset because of its capability to test an NLP model on sentiment analysis [27].To build the training and test data for our classifier, we used the publicly available datasets most frequently used by the NLP research community: SST5, Amazon Review and Yelp Review [24,25,26] A study of the latest literature in the field of textual data classification allowed us to conclude that pre-trained architectures of the Transformer type, such as BERT and some of its variants, achieve the best scores in terms of accuracy compared to classical architectures such as RNN or CNN[28].

Nine classification models were implemented by performing fine-tuning on pre-trained models based on the BERT Transformer[29] .In this study, all models were implemented using the transformer library in Python, with the appropriate tokenizer selected depending on the specific model. Both ELMo and ULMFIT employ long short-term memory (LSTM) networks. In addition to the differing operational mechanisms of these two approaches, it is worth noting that the use of transformers enables parallelization of training, which is a crucial consideration when working with large datasets. In contrast to other models, BioBERT, Bio-ClinicalBERT, and PubMedBERT have been pre-trained on medical and clinical notes, as opposed to more generic corpora such as Wikipedia and English dictionaries. Devlin et al.(2019) proposes the following range of values are recommended: Batch size: 16, 32; Learning rate (Adam): 5e-5, 3e-5, 2e-5; Number of epochs: 2, 3, 4[29].

5 Results

5.1 Dataset

The size of the corpus needed for a given task is determined by factors such as the intended use, computer processing speed, storage capacity, and the frequency and distribution of the linguistic features of diabetes in the corpus. The corpus for Indian Ocean and France contains 30,646 and 25,166 unique words , respectively,

which is considered a good representation of the dataset. The table 1 gives an insight on the content of the corpora in the two different regions.

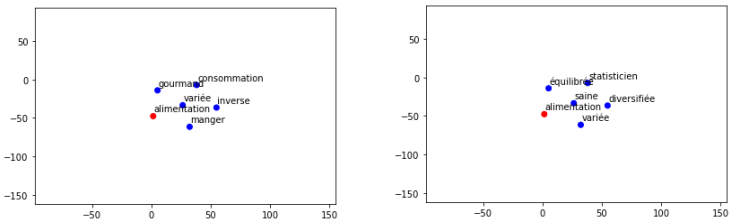
Table 1: Summary of Dataset

Variables	**Indian Ocean Region**	**France**
Number of Articles	613	607
Avg. Word/Sentence	27.8	21.8
Readability Index	14.18	12.94
Total number of words	493,744	25,166
Unique words	30,646	25,166

The vocabulary density and Readability index for Indian Ocean and France are 0.062 and 0.063 and 14.18 and 12.94, respectively. A low vocabulary density indicates complex text with many unique words, while a high ratio indicates simpler text with frequently reused words. A low density in this case also indicates that the corpus is well balanced. A readability index is an estimation of how difficult a text is to read, based on factors such as word lengths, sentence lengths, and syllable counts. The scores obtained indicate that the text is a bit difficult to read and is best suited for college graduates, with a score of 0-10 considered for professionals.

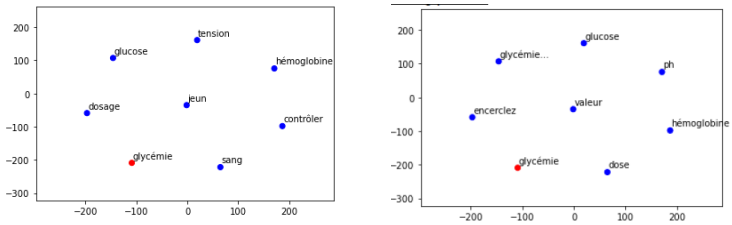
5.2 Word Association

The results of applying Word2Vec to the corpus are illustrated in figures 2, 3, 4, 5. While some words association are common to both region, for alimentation (eaiting in english) in the indian ocean region the words "diversifié","équilibré" which means diversified and balanced respectively does not appear. For the next word "glycemie", which basically mean blood sugar for france the word "tension", meaning blood pressure, is does not appear in proximity. The information shown can be further interpreted by health care professionals and communication experts. The position of surrounding words in relation to the target word is crucial as it illustrates the semantic similarities between words, with closer words being more similar. The implemented code utilizes a lexicon of words related to diabetes to investigate how the disease is perceived and characterized. The number of similar words can be adjusted for any lexicon word used in the analysis.



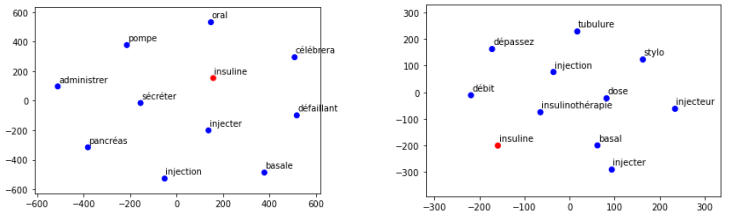
(a) Indian Ocean (b) France

Fig. 2: similar words to the word ‘alimentation’



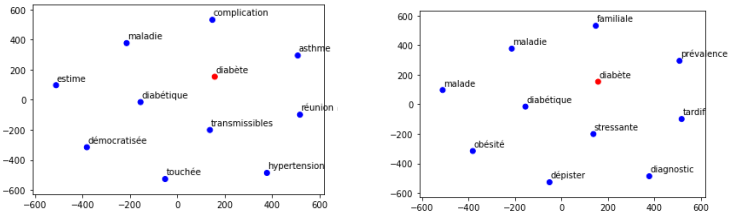
(a) Indian Ocean (b) France

Fig. 3: similar words to the words ‘Glycémie’



(a) Indian Ocean (b) France

Fig. 4: similar words to the words ‘Insuline’



(a) Indian Ocean (b) France

Fig. 5: similar words to the words 'Diabète'

5.3 Word Analogy

Table 2 shows the results for word analogies for the two corpus. This give insight to information which would not have been possible without the application of NLP. For example we can see that in indian ocean (IO)

$$regionsport + "alimentation"(eating) - hypertention$$

is equal to "manger" (eating) which in France it is equal to "endurance" which is basically more oriented to resistance to effort in sport. In IO region the

$$complication + malade(sick) - "immunitaire"(imune)$$

equals to amputation which is a major concern in Mauritius however in France it points to "neophropatie" which is linked to renal diseases. These insights can again be better interpreted by healthcare professionals. Thhis technique is a power tool in medical field.

Table 2: Word Analogy example

Analogy	**Indian Ocean**	**France**
sport + alimentation – hypertension =	manger	endurance
diabète + rénale – fruit =	cécité	insuffisance
complication + maladie – immunitaire =	amputation	néphropathie

5.4 Topic Identification

Latent Dirichlet Allocation (LDA) model was implemented to identify topics related to diabetes. The model was built with 20 different topics, where each topic is a combination of keywords, and each keyword contributes a certain weight to the topic. Table 3 provides a summary of the main subjects of the different topics around diabetes.

To evaluate the performance of the model, perplexity was calculated. Perplexity is a statistical measure that is used to determine the quality of a given topic model. A lower perplexity value indicates that the model is better at predicting the probability of the word in a given topic, hence a better model.

Table 3: Results of Topic Identification.

Topics	Indian Ocean	France
1	dépistage	glycemie-insuline
2	obésité	diabete type
3	alimentation	symptôme
4	consomation d'alcool	malade
5	patient	insuline
6	enfant	hypoglycemie
7	fruits	sucre
8	covid	alcool
9	vaccin	vitamin
10	hôpital	Glucozor
11	étude	enfant
12	sports	sommeil
13	madagascar	sports
14	alimentation	potassium
15	régime	grossesse
16	travail	chocolat
17	pharmacie	fruit
18	sucre	glycemie
19	traitement	alimentation
20	trouble	médecin

The analysis of the data from both regions revealed a high degree of similarity in the topics identified, with many common themes such as "enfant", "alimentation", "alcool", and "fruit" being present in both regions. However, there were also some notable differences, such as the presence of topics related to "grossesse"and "sommeil" in France, but not in the Indian Ocean region. These disparities represent important areas for further investigation.

In order to evaluate the performance of the Latent Dirichlet Allocation (LDA) model used in this study, model perplexity was calculated. Perplexity is a widely used measure in text analysis that provides an objective means of determining the quality of a given topic model. The perplexity values obtained for Indian Ocean region and France regions were -9.494 and -9.496, respectively.

It is worth noting that LDA is a popular tool for text analysis, providing both a predictive and latent topic representation of the corpus, however, it is equally important to identify if a trained model is objectively good or bad. The calculation of perplexity helps in determining the quality of the model. Overall, the results of this study indicate that LDA is a suitable tool for identifying similarities and differences in the topics of the data from these two regions.

5.5 Sentiment Analysis

For sentiment analysis a total of 30,000 training data and 6,000 testing data from the datasets SST5, Amazon Review, and Yelp Review were used. While there are numerous models available for this type of natural language processing

(NLP) task, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), literature suggests that models based on transformers, such as BERT, and transfer learning have been observed to exhibit superior performance in comparison to these conventional models.

Table 4: Parameters used for training sentiment analysis

Models	**Learning rate**	**Number of epochs**	**Batch size**
BERT Based	2e-5	30	4
RoBERTa-Base	2e-5	30	4
ELMo	3e-5	10	16
ULMFIT	2e-5	10	6
OpenAI GPT2	2e-5	15	6
DistilBERT	5e-5	20	4
BioBERT	5e-5	4	16
Bio-ClinicalBERT	5e-5	4	16
PubMed BERT	5e-5	4	16

The learning rate hyperparameter controls the rate or speed at which the model learns. Usually, a large learning rate enables the model to learn faster, whereas a smaller rate will take significantly longer to train but may allow the model to learn a more optimal or even globally optimal set of weights. The number of epochs is a hyperparameter that outlines the number of times that the learning algorithm will work through the entire training dataset.

Due to limitations in computational resources available on Google Colab, which was utilized as the training platform for sentiment analysis, some models that underwent a higher number of epochs with a reduced batch size exhibited comparable results to models that underwent fewer epochs but were trained using larger batch sizes.Table 4 show the parameters used to train the various models.

Table 5 presents the results of training models using a combination of data from SST5, Amazon Reviews, and Yelp Reviews. As the focus of the study is diabetes data, incorporating these three datasets in the training process offers an advantage for the final testing phase, utilizing transfer learning on the custom diabetes corpus.

Table 5: Results of Sentiment Analysis

Models	**Accuracy**	**Loss**	**Training time**
BERT Base	91.6%	0.34	3 hrs 15 mins
RoBERTa Base	**92%**	**0.27**	**03 hrs 34 mins**
DistilBERT Base	90.2%	0.43	3 hrs 42 mins
Elmo	47% early stopping	1.19	1 hr 06 mins
ULMFit	89.2%	0.89	1 hr 45 mins
OpenAI GPT2	87.4%	1.02	1 hr 23 mins
Bio-ClinicalBERT	85.9%	0.62	2 hrs 09 mins
BioMed-PubMedBER	T 88.5%	0.68	2 hrs 42 mins
BioBERT	88.6%	0.65	2 hrs 39 mins

From the analysis of nine sentiment classification models, it was observed that RoBERTa outperformed the other models with an accuracy of 92%. The model underwent training for a duration of 3 hours and 34 minutes, resulting in a loss value of 0.27. The loss metric reflects the performance of the model after each iteration of optimization. A loss value less than 0.05 would indicate underfitting, while a value greater than 0.05 would indicate overfitting. Thus, a loss value of 0.27 is considered an acceptable outcome.

Consequently, the diabetes corpus was trained using RoBERTa and the results are presented in Table 6 below:

Table 6: Results of Sentiment Analysis on Diabetes Corpus

Classes	**Indian Ocean**	**France**
Negative	294 (48%)	195 (32%)
Neutral	209 (34%)	254 (42%)
Positive	110 (18%)	158 (26%)

Table 6 reveals that a greater proportion of articles about diabetes in the Indian Ocean region are negative, at 48%, compared to those in France, which are 32% negative. This suggests that the Indian Ocean region has a higher frequency of discussion about the negative impacts of diabetes. In comparison, France has a higher proportion of neutral articles (42%), compared to positive (26%), and negative (32%), indicating that discussions in France are primarily focused on providing basic diabetes information.

6 Discussions

Health literacy refers to a set of skills necessary for effective functioning within healthcare settings, while literacy skills are becoming increasingly important for functioning within society. Low literacy has been shown to have negative effects on both health and healthcare outcomes [30]. Designing precise communication materials is one factor that contributes to improving health literacy. However, to the best of our knowledge, no such survey has been conducted in the literature for the regions of Mauritius and France.

According to the International Diabetes Federation, the prevalence of Type 2 Diabetes is 22.6% in Mauritius and 5.3% in France [4]. This disparity suggests the need for collaboration in sharing knowledge between these regions. For instance, word association surveys on diabetes reveal that the word "family" appears in the responses from France, but is completely absent in the responses from the Indian Ocean (IO) region. Similarly, the words "obesity" and "stress" are also absent in the IO region. Obesity is a significant risk factor for pre-diabetes and diabetes [31], and is a key target for the prevention and treatment of diabetes [32]. These findings highlight important factors that communication efforts should focus on, including stress management and obesity prevention.

Concerning sentiment analysis, a discernible disparity emerges when contrasting the negative, neutral, and positive sentiments associated with diabetes. Notably, the Indian Ocean (IO) region places a pronounced emphasis on highlighting the adverse aspects of diabetes. The examination of patient sentiments can facilitate issue resolution and assist decision-makers in devising adapted strategies for change, as cited in [33]. There is the need for a multidisciplinary approach i for effectively combatting the diabetes and its associated complications, which impose significant socioeconomic burden on developing countries.

7 Conclusion and future works

In conclusion, our study presents a pioneering approach utilizing Natural Language Processing (NLP) to analyze online media articles on diabetes, with implications for machine learning and healthcare research. The is a need to improve the data acquisition to include information which are shared by associations working on prevention and care of diabetes, including official communication done by public institution. On the use of NLP techniques we have demonstrated that the techniques is mastered and can be optimally applied. Our findings not only contribute to the existing body of work on NLP in the health sector but also demonstrate the potential of extending these techniques to electronic medical records and other non-medical domains. Moreover, we emphasize that combating the diabetes pandemic demands a multidisciplinary approach, encompassing prevention, education, communication, care, and management, involving not only healthcare professionals but also other stakeholders. Our study sheds light on the critical need for effective communication, education, and prevention messages delivered through online news articles, which remain a primary information source for many individuals in local and regional contexts. This research shows the importance of further efforts to ensure accurate and reliable information dissemination in the fight against diabetes.

References

1. L. J. F. Rutten, K. D. Blake, A. J. Greenberg-Worisek, S. V. Allen, R. P. Moser, and B. W. Hesse, "Online health information seeking among us adults: Measuring progress toward a healthy people 2020 objective," *Public Health Reports*, vol. 134, no. 6, pp. 617–625, 2019.
2. A. Boissonnet, M. Saeidi, V. Plachouras, and A. Vlachos, "Explainable assessment of healthcare articles with qa," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, 2022, pp. 1–9.
3. T. J. Leeper and R. Slothuus, "151How the News Media Persuades: Framing Effects and Beyond," in *The Oxford Handbook of Electoral Persuasion*. Oxford University Press, 06 2020.
4. I. D. F. D. A. 10th edition scientific committee. (2021) Idf diabetes atlas 10th edition.
5. A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet *et al.*, "Machine learning and natural language processing in mental health: systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, p. e15708, 2021.

6. E. Kang, N. Song, and H. Ju, "Contents and sentiment analysis of newspaper articles and comments on telemedicine in korea: Before and after of covid-19 outbreak," *Health Informatics Journal*, vol. 28, no. 1, p. 14604582221075549, 2022.
7. S. Wang, F. Song, Q. Qiao, Y. Liu, J. Chen, and J. Ma, "A comparative study of natural language processing algorithms based on cities changing diabetes vulnerability data," *Healthcare*, vol. 10, no. 6, p. 1119, Jun 2022. [Online]. Available: http://dx.doi.org/10.3390/healthcare10061119
8. O. Oyebode and R. Orji, "Detecting factors responsible for diabetes prevalence in nigeria using social media and machine learning," in *2019 15th International Conference on Network and Service Management (CNSM)*. IEEE, 2019, pp. 1–4.
9. ——, "Mediner: Understanding diabetes management strategies based on social media discourse," in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 1546–1553.
10. P. SV, J. M. Lorenz, R. Ittamalla, K. Dhama, C. Chakraborty, D. V. S. Kumar, and T. Mohan, "Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people's perspectives regarding covid-19 booster vaccine shots in india: Crucial to expanding vaccination coverage," *Vaccines*, vol. 10, no. 11, p. 1929, Nov 2022. [Online]. Available: http://dx.doi.org/10.3390/vaccines10111929
11. M. Canaparo, E. Ronchieri, and L. Scarso, "A natural language processing approach for analyzing covid-19 vaccination response in multi-language and geolocalized tweets," *Healthcare Analytics*, p. 100172, 2023.
12. M. S. Zulfiker, N. Kabir, A. A. Biswas, S. Zulfiker, and M. S. Uddin, "Analyzing the public sentiment on covid-19 vaccination in social media: Bangladesh context," *Array*, vol. 15, p. 100204, 2022.
13. E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn, "Diabetes on twitter: a sentiment analysis," *Journal of diabetes science and technology*, vol. 13, no. 3, pp. 439–444, 2019.
14. M. d. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodriguez-Garcia, and R. Valencia-Garcia, "Sentiment analysis on tweets about diabetes: an aspect-level approach," *Computational and mathematical methods in medicine*, vol. 2017, 2017.
15. I. De la Torre-Díez, F. J. Díaz-Pernas, and M. Antón-Rodríguez, "A content analysis of chronic diseases social groups on facebook and twitter," *Telemedicine and e-Health*, vol. 18, no. 6, pp. 404–408, 2012.
16. Y. Liu, R. Stouffs, and Y. L. Theng, "Sentiment analysis on social media for identifying public awareness of type 2 diabetes," in *The 54th International Conference of the Architectural Science Association (ANZAScA)*, 2020.
17. K. Foley, D. McNaughton, and P. Ward, "Monitoring the 'diabetes epidemic': A framing analysis of united kingdom print news 1993-2013," *PloS one*, vol. 15, no. 1, p. e0225794, 2020.
18. N. F. Syafhan, G. Chen, C. Parsons, and J. C. McElnay, "Potential of uk and us newspapers for shaping patients' knowledge and perceptions about antidiabetic medicines: a content analysis," *Journal of Pharmaceutical Policy and Practice*, vol. 15, no. 1, pp. 1–11, 2022.
19. ParseHub. (2022) The most powerful web scraper. [Online]. Available: https://www.parsehub.com/
20. Octoparse. (2019) Easy web scraping for anyone. [Online]. Available: https://www.octoparse.com/

21. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
22. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
23. ——, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
24. DeepAI. (2015) Stanford sentiment treebank dataset. [Online]. Available: https://deepai.org/dataset/stanford-sentiment-treebank
25. M. Julian. (2018) Amazon review data. [Online]. Available: https://jmcauley.ucsd.edu/data/amazon/
26. Y. Inc. (2019) Yelp open dataset. [Online]. Available: https://www.yelp.com/dataset
27. J. Wei. (2020) The stanford sentiment treebank (sst): Studying sentiment analysis using nlp. [Online]. Available: https://towardsdatascience.com/the-stanford-sentiment-treebank-sst-studying-sentiment-analysis-using-nlp-e1a4cad03065
28. S. González-Carvajal and E. C. Garrido-Merchán, "Comparing bert against traditional machine learning text classification," *arXiv preprint arXiv:2005.13012*, 2020.
29. 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.
30. N. D. Berkman, T. C. Davis, and L. McCormack, "Health literacy: what is it?" *Journal of health communication*, vol. 15, no. S2, pp. 9–19, 2010.
31. A. Boles, R. Kandimalla, and P. H. Reddy, "Dynamics of diabetes and obesity: Epidemiological perspective," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1863, no. 5, pp. 1026–1036, 2017.
32. M. S. Rendell, "Obesity and diabetes: the final frontier," *Expert Review of Endocrinology & Metabolism*, no. just-accepted, 2023.
33. L. Abualigah, H. Alfar, M. Shehab, and A. Abu Hussein, *Sentiment Analysis in Healthcare: A Brief Review*, 01 2020, pp. 129–141.

Authors

Shaad Toofanee is a Senior Lecturer at the Université des Mascareignes in Mauritius(UDM). He is presently pursuing a Phd in the field of Artificial Intelligence at the Université de Limoges France (UNILIM)in the research lab XLIM, UMR CNRS 7252. His research interests are using Mahcine Learning in the field of health and more precisely diabetes prevention, management and education. He is presently investigating the use of vision transformer for image and Transformers for Natural Language Processing. He is co-director of the programme Master in artificial Intelligence and Robotics.

Sabeena Dowlut is a Senior Lecturer at the Université des Mascareignes in Mauritius(UDM) and Head of depart of Applied Computer Science. She has a Phd in Health Litteracy from Université de la Réunion. She is also co-director of a Master programme in Health and AI which will be starting in September 2023.She is also member of Réseau francophone de littératie en santé(RÉFLIS) and a member of the Economic and Scientific Committee of AUF for the Indian Ocean and African Austral region. She is presently co-supervising a Phd thesis in the field of IoT and Health.

Nabeelah Pooloo graduated with a Bsc. Honours degree in software engineering degree from Université des Mascareignes. She was successfully selected for a fully funded government of mauritius scholarship for a Master degree in Artificial Intelligence and Robotics which was jointly offered by Université des Mascareignes and University of Limoges. She successfully completed her research internship at XLIM laboratory, UMR CNRS 7252 (University of Limoges).

Karim Tamine is an Associate Professor / Researcher at thein the research lab XLIM, UMR CNRS 7252 (University of Limoges). His research work focuses on the use of Artificial Intelligence methods in various fields such as computer graphics, security and quality of service in dynamic communication networks. He is the main resource person in artificial intelligence for the setting up of a master degree at Université des Mascareignes Mauritius. He has supervised 8 phd thesis and he is presently supervising 2 phd students.He has taken an interest in the field of application of Machine Learning in healthcare.

Damien Sauveron is Professor/ Researcher at the XLIM laboratory (University of Limoges). He is also presently dean of the faculty of science and technology.His research interests are related to Smart Card

applications and security (at hardware and software level), RFID/NFC applications and security, Mobile networks (e.g UAV fleets) applications and security, Sensors network applications and security, Smart home applications and security, Internet of Things (IoT) security, Cyber-Physical Systems security, security of Distributed Objects and Systems and security evaluation/certification processes.

A.2 DLMDISH

International Journal of Image and Graphics
(2025) 2550045 (21 pages)

DOI: 10.1142/S0219467825500457

DLMDish: Using Applied Deep Learning and Computer Vision to Automatically Classify Mauritian Dishes

Shaad Toofanee*, Omar Boudraa† and Karim Tamine‡

Department of Computer Science
XLIM, UMR CNRS 7252, University of Limoges
123 Avenue Albert Thomas, 87060 Limoges Cedex, France
**mohammud-shaad-ally.toofanee@unilim.fr*
†*omar.boudraa@gmail.com*
‡*karim.tamine@unilim.fr*

Received 25 June 2022
Accepted 14 August 2023
Published

The benefits of using an automatic dietary assessment system for accompanying diabetes patients and prediabetic persons to control the risk factor also referred to as the obesity "pandemic" are now widely proven and accepted. However, there is no universal solution as eating habits of people are dependent on context and culture. This project is the cornerstone for future works of researchers and health professionals in the field of automatic dietary assessment of Mauritian dishes. We propose a process to produce a food dataset for Mauritian dishes using the Generative Adversarial Network (GAN) and a fine Convolutional Neural Network (CNN) model for identifying Mauritian food dishes. The outputs and findings of this research can be used in the process of automatic calorie calculation and food recommendation, primarily using ubiquitous devices like mobile phones via mobile applications. Using the Adam optimizer with carefully fixed hyper-parameters, we achieved an Accuracy of 95.66% and Loss of 3.5% as concerns the recognition task.

Keywords: Deep learning; generative adversarial network; Mauritian food dataset; classification; real time mobile application.

AQ: Please provide the corresponding author details. Provide the ORCID id for all authors.

1. Introduction

The Republic of Mauritius is a small island in the Indian Ocean. The country is famous for its sandy beaches but, unfortunately, also for the high level of prevalence of diabetes and pre-diabetes. Diabetes is a "chronic condition that occurs when there are raised levels of glucose in the blood because the body cannot produce any or enough of the hormone insulin or use insulin effectively".[1] It is categorized as "not only a health crisis but a global societal catastrophe".[1] There are 22.8% of the

*Corresponding author.

Mauritian population who are diabetic, with a further 19.5% who are pre-diabetic.[a] While the existing actions by various governmental and non-governmental organizations are doing their best to fight this critical issue, the projection made by the International Diabetes Federation (IDF) for years 2030 and 2045 nonetheless shows that there will be a rise in the percentage of the population with diabetes.[2]

The above situation clearly requires an out-of-the-box, innovative approach to tackle the problem. Diabetes in Mauritius island stems from a lack of information about culinary habits. The idea is to use images of various Mauritian dishes to develop an automatic learning-based application to assist Mauritians in combating the scourge of diabetes and to assist practitioners in this task.

We propose an innovative and inclusive ecosystem with Artificial Intelligence (AI) coupled with the work undertaken by Toofanee *et al.* relating to the use of Therapeutic Patient Education and mobile applications for diabetes education and management.[3]

Based on the work of Slama-Chaudhry and Gray,[4] the WHO recommends the use of eHealth and mHealth as one of the global actions for the period 2013–2022 to prevent and control Non-Communicable Diseases (NCD). Moreover, the WHO

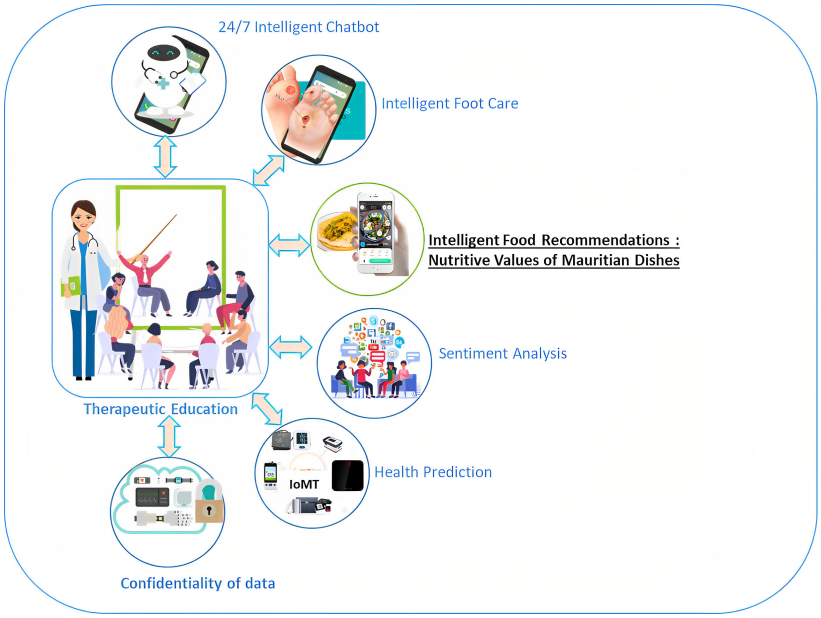


Fig. 1. Proposed overall AI-powered novel innovative ecosystem for diabetes management and education.

[a]Ministry of Health and Quality of Life, 2015.

acknowledges that AI can be used as a tool for the prevention of lifestyle diseases such as diabetes[5] as it can change how patients eventually manage their chronic conditions.[6] Figure 1 gives an overall picture of the final objective we want to achieve when dealing with the problem in a holistic view. However, this paper tackles the food recommendation part using Deep Learning (DL) when dealing with diabetes type 2 (DT2).[7] Deepat *et al.* and Foroushi *et al.* emphasized the role played by dietary and nutritional factors as a major factor in prevention and management.[8,9] While foodstuffs imported from European countries tend to give an indication of the nutritional values of the products, in Mauritius, most households are still based on homemade foods which fall outside of this category. This poses the major problem of controlling the food consumed by the people. Systems based on inputting daily food intake information manually via a mobile phone or tablet are a tedious and tiring task. Self-reporting system approaches are, however, more likely to have mistakes.[10]

AQ: Please approve the short title.

In this paper, we aim to use DL for Mauritian food recognition and also to present the first annotated dataset of Mauritian food which can be a stepping stone of an overall system for encouraging better eating habits and hence acts as primary prevention and secondary prevention of diabetes.

To the best of our knowledge, no prior research exists in the literature on the recognition of Mauritian dishes at the time of writing this paper. Therefore, we can outline our contribution as follows:

(1) Creating an annotated dataset for Mauritian dishes by overcoming lack of dataset.
(2) Proposing a DL model for Mauritian food recognition based on the above created dataset.

2. Methodology and Technology

This section provides an overview of the dataset and describes the food recommendation module architecture. Furthermore, the implementation of this network for the recognition of Mauritian dishes using a smartphone is also discussed.

While many researchers[10,13–18] emphasize on the need for quality and diversity of the dataset, there is a common understanding that automated dietary assessment will be a reality soon to assist health care professionals to curb metabolic diseases related to poor diet.

To the best of our knowledge there are two things that do not exist in relation to Mauritian dishes:

(1) A research-backed database of main Mauritian dishes.
(2) The precise nutritional values of mostly consumed Mauritian dishes.

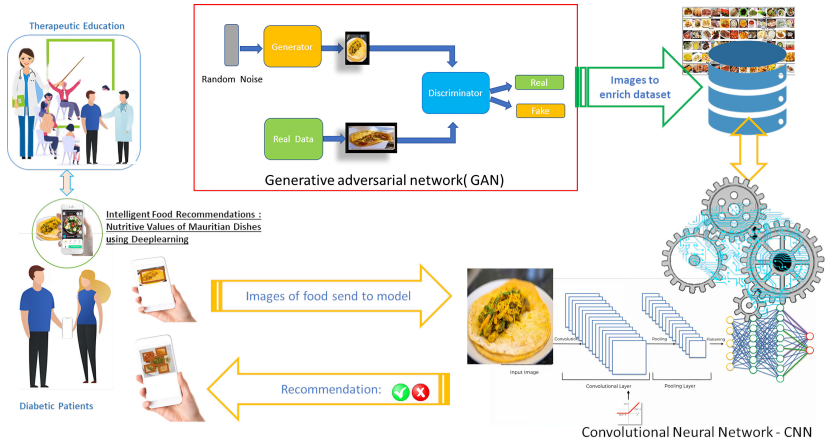


Fig. 2. Overview of our proposed architecture to help diabetic patients automatically calculate meal calories and give recommendations using GAN and CNN.

The challenge is not in the availability of mobile devices or internet connectivity, but rather in how to augment a dataset of images gathered from the internet and social networks based on the eating habits of Mauritians.

Figure 2 above expands on the overall ecosystem which was presented in the introduction section (see Fig. 1). From the available evidence, a range of dietary interventions may provide useful approaches for the management of people with type 2 diabetes, including regulation of blood glucose and lipid parameters and reduction of the risk of acute and chronic diabetic complications.[20] However, once the patient leaves the therapeutic education classes, it becomes difficult to have total control of the lifestyle. Hence, this system of automated recommendation and logging is proposed. In the long run, just scanning a dish using the mobile phone should be able to help the patient to make an informed decision, acting as a personal intelligent nutritionist for the patient.

DL models have revolutionized the state of visual object recognition and detection.[21] Most of the research papers we previewed are based on the DL architecture of Neural Networks called Convolutional Neural Networks (CNNs). In our proposed system architecture, the heart of the system is the CNN which is going to determine the content of an image as food or non-food and provide recommendations based on healthy eating habits. The classification implemented a variant of CNN which is the Mask RCNN which we explain in the following section.

CNN resembles an Artificial Neural Network (ANN) as it contains several layers. It differs from ANN since it has hidden layers which are identified as convolutional layers. The power of the CNN lies with its convolutional layers which receive input, and converts the inputs, and passes it to the next layer as shown in Fig. 3.

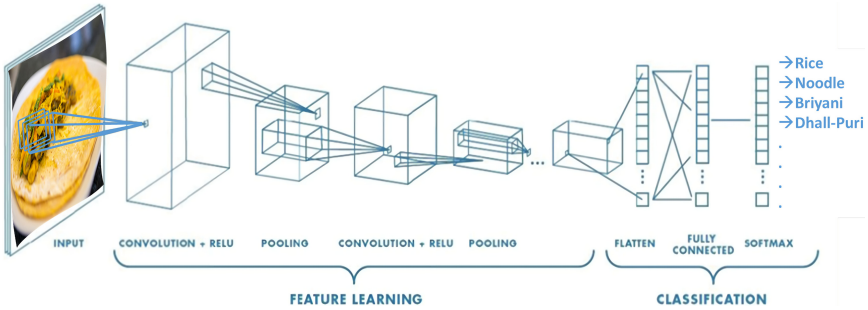


Fig. 3. Example of block diagram using CNN for food image classification.

2.1. *Classification*

CNNs are currently the most widely used DL architecture for image classification tasks. The main concept behind CNNs is that a local understanding of an image is sufficient. The principal advantage of using CNNs is that the reduced number of parameters significantly improves the learning time and decreases the amount of data required to train the system. Rather than utilizing a fully connected network of weights for each pixel, a CNN employs only the necessary weights to analyze a small patch of the image. This is the primary justification for the utilization of CNNs in the identification of Mauritian food dishes.

Although CNNs are useful for image classification, they are not ideal for image segmentation tasks. The main reason for this is that in classification tasks, we typically know the number of classes beforehand, while in object detection, we cannot determine how many regions of interest will be present in advance. In 2014, Girshick *et al.* proposed a solution to this issue, which is illustrated in the image below.[22] Girshick *et al.* proposed a method where they extract only 2000 regions from the image by a selective search algorithm, which we will not detail in this paper.[22] They called these regions, region proposals. So, instead of trying to classify a huge number of regions, we can now work with only 2000 regions. However, it clearly impacts performance as it must evaluate 2000 regions for each image. This makes it less likely to be implemented if we are working on a system that has a time limitation constraint.

Girshick, 2015 eventually proposed a new method called Fast R-CNN, which is an extension of R-CNN.[23] It is a CNN and state-of-the-art in terms of image segmentation and instance segmentation. It removed the mandatory step of always feeding 2000 region proposals to the CNN as depicted in Fig. 5 .

Faster R-CNN proposed by Ren *et al.* in 2015 with basically an alternative to the selective search algorithm was used to find region proposals as it was costly in terms of execution time and slows the network.[24]

AQ: Please provide the citation for Figure 4.

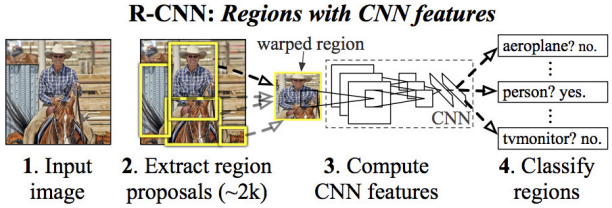


Fig. 4. R-CNN (Girshick *et al.* in 2014).[22]

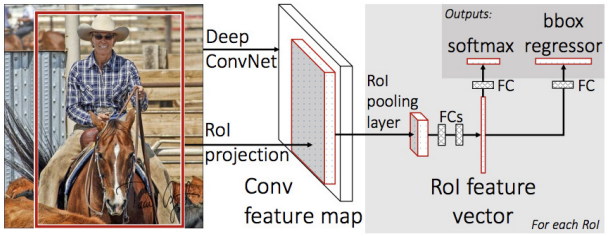


Fig. 5. Fast R-CNN.[23]

In 2017, He *et al.* presented Mask R-CNN extending Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.[25] There are two main types of image segmentation that fall under Mask R-CNN:

(1) Semantic Segmentation.
(2) Instance Segmentation.

2.2. *Enrichment*

Although Mask R-CNN was selected as the primary system for our proposed approach, the limited size of our dataset remained a significant challenge. To address this issue, we propose to employ geometric transformation and Generative Adversarial Network (GAN) techniques, which were initially introduced by Goodfellow *et al.* in 2014. Yann Lecun has described GAN as "the most interesting idea in the last 10 years in Machine Learning," and GANs are now recognized as a type of generative model.[26]

GANs have the capacity to generate new content that is comparable to the original content, which makes them a valuable tool for addressing dataset limitations. We explain this concept in greater detail in the following section. Figure 6 depicts the overall system architecture, including the type of CNN employed, the source of the dataset, and the augmentation techniques employed.

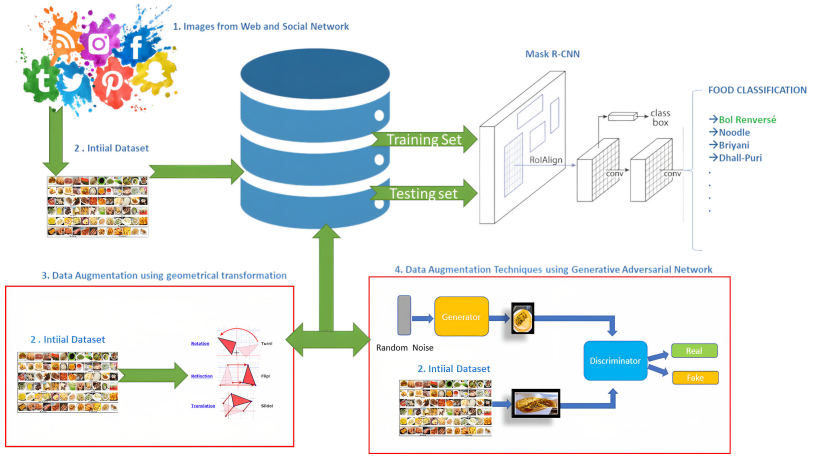


Fig. 6. Detailed block diagram of food classification.

3. Experiments and Results

This section introduces our enriched dataset and outlines the performance measures employed in our evaluation. We subsequently discuss the training protocols used, before presenting the numerical results and comparisons. Finally, we provide a summary of the devices utilized in our experiments.

3.1. *Construction of the dataset*

A dataset is essential for the training of the models for the recognition of the different parts of the dish. Since there are no past surveys that determine the most consumed dishes in Mauritius, our research for the construction of the dataset was based on data found from search engine result pages listing the most popular foods consumed by the Mauritian population. The images for the generated classes (22) were obtained by scraping image data from Google Image. A total of 6927 RGB images were obtained from the scraping process. This original dataset was carefully restricted to nine classes for the training, namely: (*Briyani/Fried noodles/Bolrenverse/Gateau Piment/French fries/Rougaille saucisse/Fried chicken/White rice/Fried egg*). For each of the experiment types, the images were manually annotated using the VGG Image Annotator tool which allows the annotations to be exported using the COCO format for the masks. Inputting the labels in this data format allows for easier processing if required.[27]

Figure 7 shows an example of images annotation. The region of interest is surrounded by the yellow line representing the polygon boundary for that region. Each image is therefore processed in the same way, that is, the masks are defined based on

Fig. 7. Example of image annotation using the VGG Image Annotator tool.

Fig. 8. Assign mask to its corresponding class based on the image's characteristics.

each image's characteristics. In that way, each mask is assigned to its corresponding class. Logically, only one class can be assigned per class (see Fig. 8). We provide the Entire Mauritian Food Project Dataset for public access in the following link available via GitHub: `https://github.com/sheik61/M2-Mauritian-Food-Project`.

3.2. *Image data augmentation*

Machine learning models require a huge amount of data for training. The more data we have, the more the model gains in performance since it will have the possibility of capturing more behaviors in the learning part without causing over-fitting which can cause the opposite effect. So before being able to test a machine learning or DL model, we must make sure that we have a rich dataset for the smooth running and training of the model. Where applicable, data augmentation methods will be used.

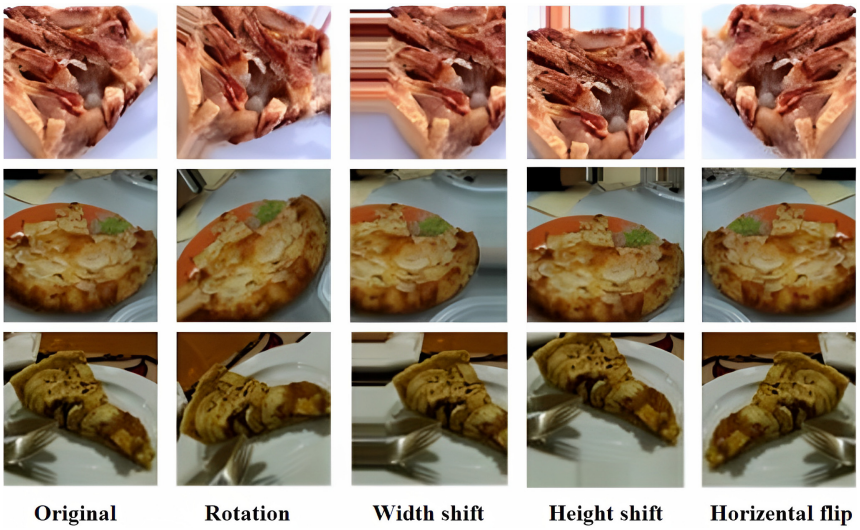


Fig. 9. Different transformations applied to an input image.

In general, data augmentation[28] is a technique that can be used to artificially increase the size of a training set by creating modified data from existing ones. It is a good practice to use augmentation if the initial dataset is too small to train, or even result in better model performance.

Data augmentation improves the model prediction accuracy. Also, it prevents data scarcity and it frames better data models.

3.2.1. *Data augmentation using transformations*

Here, concerned operations involve applying different transformations to the original images, resulting in multiple altered copies of the same image. Each copy, however, is different from the other in some ways depending on the augmentation techniques we apply like *Edge enhancement, Translation, Rotation and Flipping* . (see Fig. 9).

The application of these minor variations to the original image does not change its target class, but rather provides a new perspective on capturing the object in real life. And so, we use it quite often to build DL models. In our case, simple operations have been applied such as Flipping, Clipping, Rotation and Saturation modifying, generating consequently $473 \times 3 = 1419$ in total. Figure 10 shows some examples of transformed images.

Although this technique does not require a lot of RAM and it generates a large number of images in a short time; it does not provide images that the model does not already know, even if it learns more.

Fig. 10. Enrich original dataset using elementary operations.

3.2.2. *Data augmentation using DCGAN*

In our implementation, we used a DCGAN which is a modified version of the GAN, except that it explicitly uses convolution and convolution transposition layers in the Discriminator and Generator, respectively. It was first described by Radford *et al.*[28] The Discriminator is composed of striped convolution layers, batch norm layers and LeakyReLU activations. The input is a $3 \times 64 \times 64$ image and the output is a scalar probability that comes from the actual data distribution. The Generator is composed of convolutional transposition layers, batch norm layers, and ReLU activations. The input is a latent vector, zz, which is taken from a standard normal distribution and the output is a $3 \times 416 \times 416$ RGB image. Striped Conv-transpose layers allow the latent vector to be transformed into a volume having the same shape as an image. In this paper, the authors also give some recommendations on setting up optimizers, calculating loss functions, and initializing model weights, all of which will be explained in the following sections. Its architecture can be represented in Figure 11.

To test the DCGAN for the first time, even before we implemented the data augmentation code by transformations, we gave the model the dataset of Mauritian

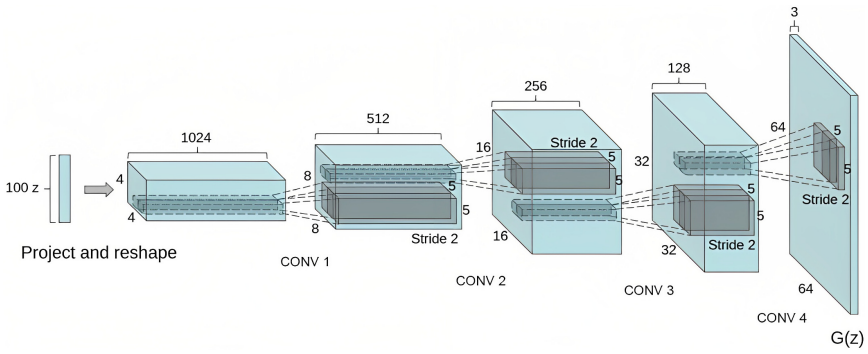


Fig. 11. Overview of DCGAN Generator.[28]

(a) (b)

(c) (d)

Fig. 12. Generated images after application of the DCGAN, (a) Bol reverse (Original), (b) Bol reverse (DCGAN on the whole dataset), (c) French fries (Original), (d) French fries (DCGAN on a selected label in the dataset).

dishes without labeling, i.e. we gave the model the 1419 images (all dishes combined) only to have an idea of the results that could be obtained.

Training 1419 images took 5 days (with one workstation) and generated unrecognizable dishes, which makes sense given that the images do not represent a single, or a unique label. Even if the images do not represent a known dish in the dataset, we can see that the DCGAN was still able to generate a plate or a bowl, which tells us that it is indeed food in these images (see Fig. 12).

After testing the DCGAN model on the entire images of the dataset, we notice that the results were not conclusive, which is understandable because no matter how efficient the model is, it will probably never be able to generate logical images

without labeling. We wanted to test the model with each dish, but as we explained, this database is very poor in terms of images.

We therefore applied transformations to labeled images to have approximately 1500 images per label and launched the DCGAN with these images plus the original images. The label "French Fries" contains 300 basic images (among the 1419 images). Then we applied transformations to them. We generated 5 images per image, so we got 1500 images for this label plus the 300 original images. We provided GAN with 1800 images, and it generated 10,000, so in total we have 11,800 images of "French Fries".

We empirically notice that by giving the algorithm precise labels, it generates better results, but also, the more data we give it, the better it generates images resembling the database.

3.3. *Evaluation criteria*

Five measures are used to evaluate our classification system efficiency and to study its robustness to perturbations (Namely: Loss, Accuracy, the Sensitivity (or the Selectivity) (SEN), the Specificity (SPEC) and the Precision (PRE)), as follows:

- **Loss:** The model error is computed periodically using the *Categorical Cross Entropy* function (also called *Softmax Loss Function*):

$$\text{Loss} = \sum_{j=1}^{m} \sum_{i=1}^{n} (y_{ij} \times \log(\hat{y}_{ij})). \tag{1}$$

  The final layer output consists of an expected class probabilities vector $\hat{y}$. The objective is to detect the best parameters minimizing the difference between the expecting $\hat{y}$ and the true y concerning the input $\hat{x}$ for a total of n trails (images in our case) among m classes.
- **Accuracy:** In reverse, this criterion designates the proportion of the correct predictions made by the model:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{2}$$

  where TP, FP, TN, FN denote, respectively, the number of true positives, false positives, true negatives and false negatives.
- **SEN:** In medicine, the Sensitivity of a diagnostic test is its ability to detect a maximum of positive results when a hypothesis is verified (i.e. to have the fewest false negatives):

$$\text{SEN} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{3}$$

- **SPEC:** Contrariwise, the Specificity is the ability to detect a maximum of negative results when a hypothesis is verified (i.e. to have the fewest false positives):

$$\text{SPEC} = \frac{\text{TN}}{(\text{TN} + \text{FP})}. \tag{4}$$

- **PRE:** For a set of examples, Precision is the ratio of the correctly classified images:

$$\text{PRE} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \tag{5}$$

3.4. *Technical details*

In this part, we discuss how we set up different hyper-parameters and protocols concerning our architecture.

3.4.1. *Procedure of data augmentation using transformations*

In order to generate new images with this method, we used the **Keras** class: **ImageDataGenerator**. **ImageDataGenerator** generates real-time data augmentation batches of tensor image data. This class provides a quick and easy way to enhance images. It offers a host of different augmentation techniques like uniforming, rotation, shifts, reversals, changing brightness, and many more. At each epoch, the **ImageDataGenerator** class ensures that the model receives new variations of the images. However, it only returns the modified photos and does not include them in the original corpus. If this were the case, the model would be exposed to the original images many times, causing the model to overfit.

3.4.2. *Generator*

The Generator uses the **Conv2DTranspose** layers to produce an image from a seed (random noise). We start with a Dense layer that takes this seed as input, then we resample several times until we reach the desired image size of $416 \times 416 \times 3$. We use the **LeakyReLU** activation for each layer, except for the output layer which uses **Tanh** to get values ranging from -1 to $+1$. However, in some middle layers we apply **Batch Normalization** to ensure faster model convergence and improve its accuracy using Mean and Standard Deviation Neurons values.

3.4.3. *Discriminator*

Discriminator is just a CNN-based image classifier. As is the case with the Generator, we make use of the **LeakyReLU** activation for each layer, except for the output layer, which uses **Sigmoid** function to get values ranging from 0 to $+1$ and **Batch Normalization** to accelerate the convergence process.

3.4.4. *Loss and optimizer*

The Discriminator's loss serves as a metric for evaluating the Discriminator's ability to differentiate between real and fake images. The loss is calculated by comparing the Discriminator's predictions for real images to an array of 1 s, and its predictions for fake (generated) images to an array of 0 s.

In contrast, the Generator loss measures the extent to which the Generator was able to deceive the Discriminator. When the Generator performs well, the Discriminator will classify fake images as real (or 1). This loss is calculated by comparing the Discriminator's decisions on the generated images to an array of 1 s.

To achieve the desired outcome and update the synaptic weights ω_i of neurons in the CNN, an optimizer algorithm is utilized to compile the entire model. Choosing an appropriate optimizer algorithm is critical, as it can have a significant impact on the quality of outcomes achieved and the associated costs. Making the right choice of optimizer algorithm is therefore essential for ensuring the success of the CNN steps in achieving the desired solution.

Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD) represent the most popular and widely-used optimizers.

SGD is a classical algorithm used often along with Momentum, for the purpose of accelerating it in the appropriate direction by accumulating the gradient of the previous steps to control the direction to track. For each term ω_t awaiting convergence:

AQ: Please approve edit

$$\begin{aligned} v_t &:= \alpha \times v_t + \eta \times \nabla \sum_{i=1}^{m} L_i\left(\omega_t\right), \\ \omega_t &:= v_t + \omega_t, \end{aligned} \tag{6}$$

where η: Learning rate, we set to 10^{-4} in our tests,
α: Coefficient of Momentum, we fix to 0.9 as recommended,
v_t: Retained gradient, initialized generally to 0,
m: Size of sample (number of elements),
L: SGD Objective function to minimize, by default, it referred to BCE loss function.

Our decision is made for new efficient **Adam** optimizer. It combines both **RMSProp** and **Momentum** optimizers and calculates adaptive learning amounts, in quick time, for each parameter ω_i till convergence,[29] as follows:

$$\begin{aligned} v_t &= \beta_1 \times v_{t-1} - (1-\beta_1) \times g_t, \\ s_t &= \beta_2 \times s_{t-1} - (1-\beta_2) \times g_t^2, \\ \Delta\omega_t &= -\eta \frac{v_t}{\sqrt{s_t + \epsilon}} \times g_t, \\ \omega_{t+1} &= \omega_t + \Delta\omega_t, \end{aligned} \tag{7}$$

where η: Initial learning rate,
g_t: Gradient at time t along ω_i,
v_t: Exponential average of gradients along ω_i,
s_t: Exponential average of square of gradients along ω_i,
β_1 and β_2: Hyper-parameters, kept by authors around 0.9 and 0.999, respectively,
ϵ is chosen to be 10^{-8}.

Adam optimizer needs little memory requirements and is more appropriate for problems with very noisy/or sparse gradients compared to SGD.

3.4.5. *Material and software*

In this study, multiple experiments were conducted on three high-performance workstations featuring Nvidia Gtx GPUs and substantial RAM capacities ranging from 16 to 24 GB. In addition, Google Colaboratory, an online platform providing free access to GPU hardware acceleration and the Jupyter editor, was utilized for DL projects. The Tensor Processing Unit (TPU), an integrated circuit developed by Google for the purpose of accelerating AI systems, was also employed. The primary libraries utilized in this study included:

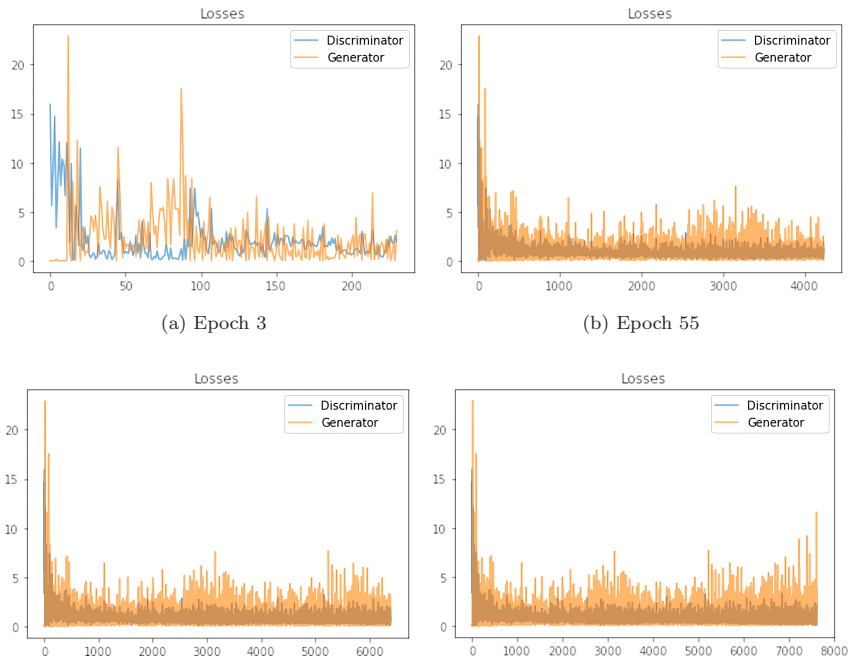
- **Roboflow:** It is a computer vision development framework for improving data collection, preprocessing, and model training techniques. Roboflow has public datasets easily accessible to users and allows them to upload their own custom data. Roboflow accepts different annotation formats for labeling including COCO JSON and CSV. Data pre-processing includes steps such as image deskewing, and resizing, contrast enhancement and data augmentation.
- **TensorFlow:** Developed by Google researchers, TensorFlow is an open-source tool for machine learning, DL, statistical and predictive analytics. Like similar platforms, it aims to streamline the development and execution of advanced analytical applications for data scientists, statisticians, and modelers.
- **Imageio:** Imageio is a Python library that provides an easy interface for reading and writing a wide range of image data, including animated images, volumetric data, and scientific formats.
- **OpenCV:** OpenCV is a well-known open-source computer vision, machine learning, and image processing library that plays a major role in the real-time operations that are essential to today's systems. It can be used to process images and videos in order to identify objects, faces, and even handwriting documents.

3.5. *Results*

By varying the quantity of images to be injected into the Generator and by following in real time the evolution of the DCGAN loss. Figure 13 gives a numerical overview of the obtained results. We notice that by giving the algorithm precise labels, it generates better results, but also, the more data we give it, the better it generates images resembling the database. Moreover, it should be noted that the average time required for each of the above-mentioned experiments is equal to 1454 s. After generating new images of the dataset, we test our data in a CNN model, with the goal being to assess Accuracy and Loss for the classification issue. The full CNN model contains 2,081,234 parameters (weights and biases).

AQ: The sentence "By varying.." seems incomplete. Kindly check for continuity.

We fed the model 12,000 images with three different labels as input. In addition, we separated the training data from the test data (80% to 20% proportion). We launched the optimizer for 100 epochs with a batch size of 128, the results of pre-mentioned indices at the end of the process are given as follow: Accuracy = 95.66%/Loss = 3.5%/SEN = 96.74%/SPEC = 94.58% and PRE = 96.88%.



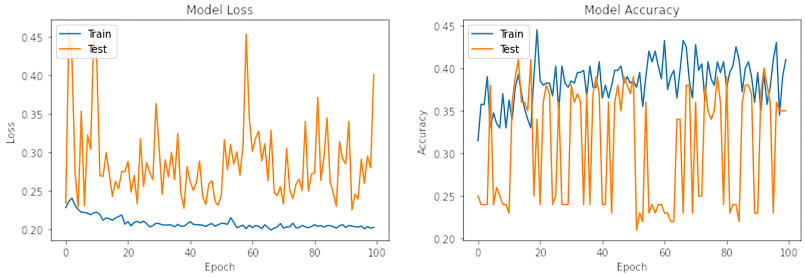
(a) Epoch 3 (b) Epoch 55

(c) Epoch 83 (d) Epoch 99

Fig. 13. Graphs highlighting Discriminator and Generator losses in different epochs.

The results obtained in this study provide strong evidence to support the efficacy, precision, and robustness of our proposed method. A visual representation of the (Loss/Accuracy) curve over the training process, as depicted in Fig. 14, was generated using a randomly selected subset comprising ($\tilde{5}\%$) of the total dataset. Finally, we find that our dishes' identifying architecture is performing well and achieving an excellent rate, which allowed us to validate it. A synthetic state-of-the-art method comparison is presented in Table 1 where our proposed method obtains the best evaluation metric values.

The table shows that the architecture proposed and tested on a restricted number of images which were augmented using several data augmentation techniques performed relatively well as compared to food classification based on other food databases. The algorithms experimented on Mauritian food dataset managed an Accuracy of 95.66% as opposed to the accuracy achieved by other works carried out on different food datasets. However, compared to the work of Termritthikun *et al.*[11] concerning food recognition using smartphones and DL on a Thai food dataset, they concentrated in reducing the processing time also. It should be noted that GAN is time consuming and processing time was not the main guiding criteria.



(a) (b)

Fig. 14. Curves highlighting Loss and Accuracy variation during 100 epochs.

Table 1. Summary of classification results achieved by other six reputed methods applied on various popular food dataset images.

Rank	Model	Authors	Datasets	Accuracy (%)	Loss (%)
1	**DLMDish**	**Proposed**	**Mauritian food**	**95.66**	**3.5**
2	DeepFood	Liu *et al.*[13]	Food-101	93.7	13
3	NU-InNet 1.0	Termritthikun *et al.*[11]	Thai food	92.3	18.16
4	MTCNet	Lu *et al.*[19]	NIAD	>91	<20
5	NutriNet	Mezgec *et al.*[10]	Fake-food images	88.6	30
6	EfficientNetB2	Konstantakopoulos *et al.*[15]	MedGRFood	83.4	6.5
7	Inception-V3	Ma *et al.*[16]	ChinaMartFood-109	78.26	19.79

In terms of creating of the dataset, they were collected by scraping images from the internet and popular Facebook pages, ideally this should be done with the help of a specialist nutritionist who has a good representation of the eating practices of the Mauritian multi-cultural population. Termritthikun *et al.* dataset was also collected from Google, Bing and Flickr.[11] Dataset used by Konstantakopoulois *et al.* was also taken from the web directly and they used image recognition techniques based on the CNN, transfer learning, data augmentation and fine-tuning techniques to achieve an acceptable accuracy.[15]

From various sources it can be noted that there is no one general purpose food database, but several datasets were set-up based on the countries and region. Liu *et al.*[13] worked on Asian food dataset and used and applied K-fold cross optimization with K set to 5 and the training and testing sets split as 75% to 25%. Unlike this study, their research extended to determining portion size, which poses a significant challenge and can be explored in future work.

The latest work by Jiji and Rajesh,[14] Ma *et al.*[16] worked on calculating nutritional values using images of food. It should be noted that there is no precise nutritional values of dishes for Mauritian food. However, if this is made available, this work can be escalated to take this information in consideration and provide calorie values of food images which are processed.

Many prospects may be cited, to enrich and complete our study. Among them:

- Provide a positive or negative recommendation after recognizing the input dish image and approximatively calculate the Nutrition calories value (mainly for: Protein (cal/100 g), Fat (cal/100 g), Carbohydrate (cal/100 g), Fiber (mg/100 g), Vitamin C (mg/100 g), Calcium (mg/100 g) and Iron (mg/100 g)).
- Improve the noisy images by using another generation of GANs which is called Variational AutoEncoder (VAE). It mainly involves two elements (Encoder followed by Decoder); and it is used to reconstruct the preliminary results. Unlike a traditional autoencoder, which maps the input to a latent vector, a VAE maps the input data into the parameters of a probability distribution, such as the mean and variance of a Gaussian distribution.[30] The whole process is manually controlled by an expert which decides whether the input image needs or not this post-processing step.
- Make improvements to the classification procedure so that each plate will be extracted separately.
- Consider an additional procedure that effectively addresses the overlapping of two close dishes upon detection.
- Context feature extraction based on attention mechanism when dealing with image segmentation can also be investigated when separating different elements that constitute a Mauritian plate.[31]
- Find an automatic adaptation of the CNN and GAN network parameters to the characteristics of the input image.

4. Conclusion

Despite the availability of several food image datasets, it should be noted that none are adapted to Mauritian dishes primarily because of the difference in eating habits. The work presented in this paper and the promise in terms of efficiency of other DL architectures for image classification and object detection means that there are still more challenges that can be undertaken in this field and to produce state-of-the-art results which can enable the creation of a full dietary food recognition and recommendation system. In this paper, we have been able to apply DL on a set of Mauritian dish for the purpose of classification by using data augmentation techniques including the GAN.

References

1. M. East and N. Africa, "IDF diabetes atlas," *Diabetes* **20**, 79 (2017).
2. H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. N. Chan, J. C. Mbanya, M. E. Pavkov, A. Ramachandaran, S. H. Wild, S. James, W. H. Herman, P. Zhang, C. Bommer, S. Kuo, E. J. Boyko and D. J. Magliano, "IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.* **183**, 109119 (2022).

3. S. A. Toofanee, B. S. Dowlut, M. Balcou-Debussche, X. Debussche, V. Lahausse and L. Nisa, "A mobile application to empower diabetic patients enrolled in a therapeutic patient education programme in Mauritius," in *IST-Africa 2022 Conf. Proc.*, eds. M. Cunningham and P. Cunningham (IST-Africa Institute and IIMC, 2022).
4. A. Slama-Chaudhry and A. Golay, "PRACTICE Patient education and self-management support for chronic disease: Methodology for implementing patient-tailored therapeutic programmes," *Public Health Panorama* **5**(2–3), 357–361 (2019).
5. J. Chaki, S. Thillai Ganesh S. K. Cidham and S. A. Theertan, "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review," *J. King. Saud Univ. Comput. Inf. Sci.* **34**(6), 3204–3225 (2020), doi:10.1016/j.jksuci.2020.06.013.
6. E. Topol, The Topol Review. Preparing the healthcare workforce to deliver the digital future, National Health Service (2019), pp. 1–48, `https://topol.hee.nhs.uk/`.
7. J. Mattei, V. Malik, N. M. Wedick, F. B. Hu, D. Spiegelman, W. C. Willett and H. Campos, "Reducing the global burden of type 2 diabetes by improving the quality of staple foods: The Global Nutrition and Epidemiologic Transition Initiative," *Global. Health* **11**, 23 (2015), doi:10.1186/s12992-015-0109-9.
8. M. Deepa, R. M. Anjana and V. Mohan, "Role of lifestyle factors in the epidemic of diabetes: Lessons learnt from India," *Eur. J. Clin. Nutr.* **71**(7), 825–831 (2017), doi:10.1038/ejcn.2017.19.
9. N. G. Forouhi, A. Misra, V. Mohan, R. Taylor and W. Yancy, "Dietary and nutritional approaches for prevention and management of type 2 diabetes," *BMJ* **361**, k2234 (2018), doi:10.1136/bmj.k2234.
10. S. Mezgec and B. Koroušić Seljak, "Deep neural networks for image-based dietary assessment," *J. Vis. Exp.* **169**, e61906 (2021), doi:10.3791/61906.
11. C. Termritthikun, P. Muneesawang and S. Kanprachar, "NU-InNet: Thai food image recognition using convolutional neural networks on smartphone," *J. Telecommun. Electron. Comput. Eng.* **9**(2–6), 63–67 (2017).
12. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1–9, doi:10.1109/CVPR.2015.7298594.
13. C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkaraneand and Y. Ma, "DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment," in *Int. Conf. Smart Homes and Health Telematics* (Springer, Cham, 2016), pp. 37–48.
14. G. W. Jiji and A. Rajesh, "Food sustenance estimation using food image," *Int. J. Image Graph.* **20**(04), 2050034 (2020).
15. F. S. Konstantakopoulos, E. I. Georga and D. I. Fotiadis, "Mediterranean food image recognition using deep convolutional networks," in *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE, 2021), pp. 1740–1743.
16. P. Ma, C. P. Lau, N. Yu, A. Li, P. Liu, Q. Wang and J. Sheng, "Image-based nutrient estimation for Chinese dishes using deep learning," *Food Res. Int.* **147**, 110437 (2021).
17. Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (2021), pp. 8903–8911.
18. W. Wang, W. Min, T. Li, X. Dong, H. Li and S. Jiang, "A review on vision-based analysis for automatic dietary assessment," *Trends Food Sci. Technol.* **122**, 223–237 (2022).
19. Y. Lu, T. Stathopoulou, M. F. Vasiloglou, S. Christodoulidis, Z. Stanga and S. Mougiakakou, "An artificial intelligence-based system to assess nutrient intake for hospitalised patients," *IEEE Trans. Multimed.* **23**, 1136–1147 (2020).

AQ: Please provide the page number for Ref. 3.

AQ: Please provide the citation for Ref. 12.

20. O. Ojo, "Dietary intake and type 2 diabetes," *Nutrients* **11**(9), 2177 (2019).
21. Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
22. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2014), pp. 580–587.
23. R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Computer Vision* (2015), pp. 1440–1448.
24. S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, Vol. 28, (2015).
25. K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Computer Vision* (2017), pp. 2961–2969.
26. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, Vol. 27 (2014).
27. A. Dutta, A. Gupta and A. Zissermann, VGG image annotator (VIA) (2016), `https:www.robots.ox.ac.uk/vgg/software/via/`.
28. A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434.
29. D. Kingsma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 2th Int. Conf. Learning Representations*, New York, USA, 2014, `https://arxiv.org/abs/1412.6980`.
30. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," (2016).
31. S. Ding, H. Wang, H. Lu, M. Nappi and S. Wan, "Two path gland segmentation algorithm of colon pathological image based on local semantic guidance," *IEEE J. Biomed. Health Inform.* **27**(4), 1701–1708 (2022).

AQ: Please provide the page number for Refs. 24 and 26.

AQ: Please provide the complete details for Ref. 30.

Shaad Toofanee is a Senior Lecturer at the University of Mascareignes in Mauritius (UDM). He is presently pursuing a Ph.D. in the field of AI at the University of Limoges (UNILIM), France. His research interests are using machine learning in the field of health and more precisely diabetes prevention, management and education.

Omar Boudraa is a Doctor on Distributed and Mobile Computing, a Temporary Research Assistant and Assistant Lecturer at Computer Science Department, University of Limoges, Limoges, France. Also, he was a visiting researcher at LIB Laboratory of Burgundy University, Dijon, France. His teaching and research interests are in image processing, AI, networks and systems administration.

Karim Tamine is a Senior Lecturer/Researcher at the XLIM laboratory (University of Limoges). His research work focuses on the use of AI methods in various fields such as computer graphics, security and quality of service in dynamic communication networks. He has supervised eight Ph.D. thesis and he is currently supervising two Ph.D. students.

B

Bibliography

Summary

Références

[1] M. E. Singer, K. A. Dorrance, M. M. Oxenreiter, K. R. Yan, and K. L. Close, "The type 2 diabetes 'modern preventable pandemic'and replicable lessons from the covid-19 crisis," *Preventive Medicine Reports*, vol. 25, p. 101 636, 2022.

[2] I. D. Federation, "Idf diabetes atlas 10th edition," World Health Organisation, Mauritius, Tech. Rep., 2021. [Online]. Available: https://diabetesatlas.org/atlas/tenth-edition/.

[3] WHO, "Biennial report 2020/2021," World Health Organisation, Mauritius, Tech. Rep., 2021. [Online]. Available: https://www.afro.who.int/sites/default/files/2022-08/WCO%20Mauritius_Annual%20Rprt_2020%202021_Web.pdf.

[4] MOH. "Planning and finance." (), [Online]. Available: https://health.govmu.org/Pages/Departments-Hospitals/Planning.aspx#:~:text=Public%20health%20services%20in%20Mauritius,around%207%25%20of%20Government%20expenditure. (accessed: 09.09.2022).

[5] MOH, "Annual report on performance for financial year 2021-2022," Ministry of Health and Wellness, Mauritius, Tech. Rep., 2022. [Online]. Available: https://health.govmu.org/Documents/Legislations/Documents/Annual%20Report%20on%20Performance%20FY%202021-2022.pdf.

[6] M. R. Rooney, M. Fang, K. Ogurtsova, *et al.*, "Global prevalence of prediabetes," *Diabetes Care*, p. dc222376, 2023.

[7] D. S. Fong, L. Aiello, T. W. Gardner, *et al.*, "Retinopathy in diabetes," *Diabetes care*, vol. 27, no. suppl_1, s84–s87, 2004.

[8] M. Balcou-Debussche, *Une approche ethnosociologique de l'éducation thérapeutique du patient dans le diabète de type 2*, 2010.

[9] K. Aggarwal, M. M. Mijwil, A.-H. Al–Mistarehi, *et al.*, "Has the future started? the current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 1, pp. 115–123, 2022.

[10] M. Woschank, E. Rauch, and H. Zsifkovits, "A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics," *Sustainability*, vol. 12, no. 9, p. 3760, 2020.

[11] K. Sharifani and M. Amini, "Machine learning and deep learning a review of methods and applications," *World Information Technology and Engineering Journal*, vol. 10, no. 07, pp. 3897–3904, 2023.

[12] M. F. Aslan and K. Sabanci, "A novel proposal for deep learning-based diabetes prediction converting clinical data to image data," *Diagnostics*, vol. 13, no. 4, p. 796, 2023.

[13] H. H. Aghdam, E. J. Heravi, *et al.*, "Guide to convolutional neural networks," *New York, NY: Springer*, vol. 10, no. 978-973, p. 51, 2017.

[14] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: progress in machine intelligence for rational drug discovery," *Drug discovery today*, vol. 22, no. 11, pp. 1680–1685, 2017.

[17] F. Jiang, Y. Jiang, H. Zhi, *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[18] P. Kumar, S. Chauhan, and L. K. Awasthi, "Artificial intelligence in healthcare: review, ethics, trust challenges and future research directions," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105 894, 2023.

[19] B. Wahl, A. Cossy-Gantner, S. Germann, and N. R. Schwalbe, "Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings?" *BMJ global health*, vol. 3, no. 4, e000798, 2018.

[20] WHO, "Ethics and governance of artificial intelligence for health: who guidance executive summary," World Health Organisation, Jun. 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240037403.

[21] M. Y. Shaheen, "Applications of artificial intelligence (ai) in healthcare: a review," *ScienceOpen Preprints*, 2021.

[22] L. Wang, Y. Zhang, D. Wang, *et al.*, "Artificial intelligence for covid-19: a systematic review," *Frontiers in medicine*, vol. 8, p. 1457, 2021.

[23] Y. Zhou, F. Wang, J. Tang, R. Nussinov, and F. Cheng, "Artificial intelligence in covid-19 drug repurposing," *The Lancet Digital Health*, vol. 2, no. 12, e667–e676, 2020.

[24] T. L. D. Health, "Artificial intelligence for covid-19: saviour or saboteur?" *The Lancet. Digital Health*, vol. 3, no. 1, e1, 2021.

[25] G. Fagherazzi and P. Ravaud, "Digital diabetes: perspectives for diabetes prevention, management and research," *Diabetes & metabolism*, vol. 45, no. 4, pp. 322–329, 2019.

[26] D. Zeevi, T. Korem, N. Zmora, *et al.*, "Personalized nutrition by prediction of glycemic responses," *Cell*, vol. 163, no. 5, pp. 1079–1094, 2015.

[27] D. McDonald, G. Glusman, and N. D. Price, "Personalized nutrition through big data," *Nature biotechnology*, vol. 34, no. 2, pp. 152–154, 2016.

[28] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[29] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *Jama*, vol. 318, no. 22, pp. 2211–2223, 2017.

[30] O. T. Kee, H. Harun, N. Mustafa, *et al.*, "Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review," *Cardiovascular Diabetology*, vol. 22, no. 1, p. 13, 2023.

[31] M. S. Ali, M. K. Islam, A. A. Das, D. Duranta, M. Haque, M. H. Rahman, *et al.*, "A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: machine learning insights," *BioMed Research International*, vol. 2023, 2023.

[32] M. Al–Tawil, B. A. Mahafzah, A. Al Tawil, and I. Aljarah, "Bio–inspired machine learning approach to type 2 diabetes detection," *Symmetry*, vol. 15, no. 3, p. 764, 2023.

[33] F. A. Khaleel and A. M. Al–Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3200–3203, 2023.

[34] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, H. H. R. Sherazi, *et al.*, "Machine learning based diabetes classification and prediction for healthcare applications," *Journal of healthcare engineering*, vol. 2021, 2021.

[35] A. Väänänen, K. Haataja, K. Vehviläinen-Julkunen, and P. Toivanen, "Ai in healthcare: a narrative review," *F1000Research*, vol. 10, p. 6, 2021.

[36] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes a systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2020.

[37] S. Bhandari, S. Pathak, and S. A. Jain, "A literature review of early–stage diabetic retinopathy detection using deep learning and evolutionary computing techniques," *Archives of Computational Methods in Engineering*, vol. 30, no. 2, pp. 799–810, 2023.

[38] A. Sebastian, O. Elharrouss, S. Al–Maadeed, and N. Almaadeed, "A survey on deep learning based diabetic retinopathy classification," *Diagnostics*, vol. 13, no. 3, p. 345, 2023.

[39] D. Das, S. K. Biswas, and S. Bandyopadhyay, "A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning," *Multimedia Tools and Applications*, vol. 81, no. 18, pp. 25 613–25 655, 2022.

[40] A. Ahne, V. Khetan, X. Tannier, *et al.*, "Extraction of explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets from 2017 to 2021: deep learning approach," *JMIR Medical Informatics*, vol. 10, no. 7, e37201, 2022.

[41] V. Anoop, "Sentiment classification of diabetes-related tweets using transformer-based deep learning approach," in *International Conference on Advances in Computing and Data Sciences*, Springer, 2023, pp. 203–214.

[42] V. Vidyadharan, M. Hamdan, and A. M. Zalzala, "An evidence–based study of diabetes prevention and management with nlp and deep learning," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2021, pp. 1–8.

[43] Z. Yu, X. Yang, G. L. Sweeting, *et al.*, "Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods," *BMC Medical Informatics and Decision Making*, vol. 22, no. 3, pp. 1–9, 2022.

[44] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[45] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta– a robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[46] M. Balcou-Debussche, V. La Hausse, M. Roddier, *et al.*, "Strengthening health literacy through structured sessions for non-communicable diseases in low-resource settings: the learning nest model," *Community Health Equity Research & Policy*, p. 2752535X231184346, 2023.

[47] K. Alexiadou and J. Doupis, "Management of diabetic foot ulcers," *Diabetes Therapy*, vol. 3, pp. 1–15, 2012.

[48] W. J. Jeffcoate and K. G. Harding, "Diabetic foot ulcers," *The lancet*, vol. 361, no. 9368, pp. 1545–1551, 2003.

[49] P. R. Cavanagh, B. A. Lipsky, A. W. Bradbury, and G. Botek, "Treatment for diabetic foot ulcers," *The Lancet*, vol. 366, no. 9498, pp. 1725–1735, 2005.

[50] I. Bartoletti, "Ai in healthcare: ethical and privacy challenges," in *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*, Springer, 2019, pp. 7–10.

[51] V. N. Shah and S. K. Garg, "Managing diabetes in the digital age," *Clinical Diabetes and Endocrinology*, vol. 1, pp. 1–7, 2015.

[52] J. S. Winter, "Ai in healthcare: data governance challenges," *Journal of hospital management and health policy*, vol. 5, no. 8, 2021.

[53] T. Schachner, R. Keller, and F. v Wangenheim, "Artificial intelligence–based conversational agents for chronic conditions: systematic literature review," *Journal of medical Internet research*, vol. 22, no. 9, e20701, 2020.

[54] A. S. Miner, L. Laranjo, and A. B. Kocaballi, "Chatbots in the fight against the covid-19 pandemic," *NPJ digital medicine*, vol. 3, no. 1, p. 65, 2020.

[55] A. Bin Sawad, B. Narayan, A. Alnefaie, *et al.*, "A systematic review on healthcare artificial intelligent conversational agents for chronic conditions," *Sensors*, vol. 22, no. 7, p. 2625, 2022.

[56] M. S. A. Toofanee, S. Dowlut, M. Hamroun, *et al.*, "Dfu–siam a novel diabetic foot ulcer classification with deep learning," *IEEE Access*, vol. 11, pp. 98 315–98 332, 2023.

[57] M. S. A. Toofanee, S. Dowlut, M. Hamroun, *et al.*, "Dfu–helper: an innovative framework for longitudinal diabetic foot ulcer diseases evaluation using deep learning," *Applied Sciences*, vol. 13, no. 18, p. 10 310, 2023.

[58] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare transforming the practice of medicine," *Future healthcare journal*, vol. 8, no. 2, e188, 2021.

[59] J. Yang, Y. Li, Q. Liu, *et al.*, "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 57–69, 2020.

[60] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare promise and potential," *Health information science and systems*, vol. 2, pp. 1–10, 2014.

[61] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annual review of public health*, vol. 36, pp. 345–359, 2015.

[62] T. Panch, H. Mattie, and L. A. Celi, "The "inconvenient truth" about ai in healthcare," *NPJ digital medicine*, vol. 2, no. 1, p. 77, 2019.

[63] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and ai for health care a call for open science," *Patterns*, vol. 2, no. 10, 2021.

[64] K. Murphy, E. Di Ruggiero, R. Upshur, *et al.*, "Artificial intelligence for good health a scoping review of the ethics literature," *BMC medical ethics*, vol. 22, no. 1, pp. 1–17, 2021.

[65] P. Solanki, J. Grundy, and W. Hussain, "Operationalising ethics in artificial intelligence for healthcare: a framework for ai developers," *AI and Ethics*, vol. 3, no. 1, pp. 223–240, 2023.

[66] WHO, "Ethics and governance of artificial intelligence for health," "World Health Organisation, Tech. Rep., Jun. 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240029200.

[67] L. Munn, "The uselessness of ai ethics," *AI and Ethics*, vol. 3, no. 3, pp. 869–877, 2023.

[68] C. Thapa and S. Camtepe, "Precision health data requirements, challenges and existing techniques for data security and privacy," *Computers in biology and medicine*, vol. 129, p. 104 130, 2021.

[69] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *arXiv preprint arXiv:1810.00184*, 2018.

[70] A. Albahri, A. M. Duhaim, M. A. Fadhel, *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion," *Information Fusion*, 2023.

[71] A. B. Arrieta, N. Díaz–Rodríguez, J. Del Ser, *et al.*, "Explainable artificial intelligence concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[72] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor xai an ontology-based approach to black–box sequential data classification explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 629–639.

[73] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, p. 107 161, 2022.

[74] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.

[75] O. Iparraguirre–Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas–Carbonell, "Application of machine learning models for early detection and accurate classification of type 2 diabetes," *Diagnostics*, vol. 13, no. 14, p. 2383, 2023.

[76] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: an empirical study," *Sensors*, vol. 22, no. 14, p. 5247, 2022.

[77] P. K. Jena, B. Khuntia, C. Palai, M. Nayak, T. K. Mishra, and S. N. Mohanty, "A novel approach for diabetic retinopathy screening using asymmetric deep learning features," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 25, 2023.

[78] G. Alwakid, W. Gouda, and M. Humayun, "Deep learning–based prediction of diabetic retinopathy using clahe and esrgan for enhancement," in *Healthcare*, MDPI, vol. 11, 2023, p. 863.

[79] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Diabetic foot ulcer detection using deep learning approaches," *Sensors International*, vol. 4, p. 100 210, 2023.

[80] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O'Shea, D. Gillespie, and N. D. Reeves, "Analysis towards classification of infection and ischaemia of diabetic foot ulcers," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2021, pp. 1–4.

[81] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[82] J. McCarthy *et al.*, "What is artificial intelligence," 2007.

[83] R. Khalkar, A. Dikhit, and A. Goel, "Handwritten text recognition using deep learning (cnn and rnn)," *IARJSET*, vol. 8, pp. 870–881, Jun. 2021.

[84] H. Alkabbani, A. Ahmadian, Q. Zhu, and A. Elkamel, "Machine learning and metaheuristic methods for renewable power forecasting: a recent review," *Frontiers in Chemical Engineering*, vol. 3, p. 665 415, Apr. 2021.

[85] Y. Zhang, X. Jiang, and S.-H. Wang, "Fingerspelling recognition by 12-layer cnn with stochastic pooling," *Mobile Networks and Applications*, Feb. 2022.

[86] X. Chen, R. Fu, Q. Shao, *et al.*, "Application of artificial intelligence to pancreatic adenocarcinoma," *Frontiers in Oncology*, vol. 12, p. 960 056, 2022.

[87] R. Chatterjee, "Fundamental concepts of artificial intelligence and its applications," *Journal of Mathematical Problems, Equations and Statistics, 1 (2)*, pp. 13–24, 2020.

[88] S. V. Mahadevkar, B. Khemani, S. Patil, *et al.*, "A review on machine learning styles in computer vision—techniques and future directions," *IEEE Access*, vol. 10, pp. 107 293–107 329, 2022.

[89] S. Fahle, C. Prinz, and B. Kuhlenkötter, "Systematic review on machine learning (ml) methods for manufacturing processes–identifying artificial intelligence (ai) methods for field application," *Procedia CIRP*, vol. 93, pp. 413–418, 2020.

[90] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi, *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[91] M. M. Taye, "Understanding of machine learning with deep learning: architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2073-431X/12/5/91.

[92] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[93] R. Saravanan and P. Sujatha, "A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 945–949.

[94] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to reinforcement learning," *Deep reinforcement learning: fundamentals, research and applications*, pp. 47–123, 2020.

[95] W. Zhang, H. Li, L. Yongqin, H. Liu, Y. Chen, and X. Ding, "Application of deep learning algorithms in geotechnical engineering: a short critical review," *Artificial Intelligence Review*, vol. 54, pp. 1–41, Dec. 2021.

[96] R. Dwivedi, D. Dave, H. Naik, *et al.*, "Explainable ai (xai): core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[97] Fidle, *Formation introduction au deep learning*, https://fidle.cnrs.fr/supports, [Accessed 11-09-2023], 2023.

[98] L. Alzubaidi, J. Zhang, A. J. Humaidi, *et al.*, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[99] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[100] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[101] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[102] 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*, 2017, pp. 4700–4708.

[103] D. Batista, *Convolutional Neural Networks for Text Classification — davidsbatista.net*, https://www.davidsbatista.net/blog/2018/03/31/SentenceClassificationConvNets/, [Accessed 03-08-2023], 2021.

[104] S. Chu, *Computer vision - samuel's' blog*, https://samueljchu.com/showcase/A.I./Computer_Vision/, [Accessed 03-08-2023].

[105] M. M. Taye, "Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions," *Computation*, vol. 11, no. 3, p. 52, 2023.

[106] A. Rosebrock, *Convolutional neural networks cnns) and layer types*, https://pyimagesearch.com/2021/05/14/convolutional-neural-networks-cnns-and-layer-types/, [Accessed 03-08-2023], 2021.

[107] S. Raziani and M. Azimbagirad, "Deep cnn hyperparameter optimization algorithms for sensor-based human activity recognition," *Neuroscience Informatics*, vol. 2, no. 3, p. 100 078, 2022.

[108] B. Khuong, *He basics of recurrent neural networks (rnns)*, https://pub.towardsai.net/whirlwind-tour-of-rnns-a11effb7808f, [Accessed 04-08-2023], 2021.

[109] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[110] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (lstm) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, 2019.

[111] N. A. Smith, "Contextual word representations putting words into computers," *Communications of the ACM*, vol. 63, no. 6, pp. 66–74, 2020.

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

[113] P. J. Worth, "Word embeddings and semantic spaces in natural language processing," *International Journal of Intelligence Science*, vol. 13, no. 1, pp. 1–21, 2023.

[114] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[115] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[116] K. Doshi, *Transformers explained visually*, https://towardsdatascience.com/transformers-explained-visually-not-just-how-but-why-they-work-so-well-d840bd61a9d3, [Accessed 27-09-2023], 2021.

[117] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[118] K. Doshi, *Transformers explained visually part 1*, https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452, [Accessed 27-09-2023], 2020.

[119] A. Soleimany, *Recurrent neural networks and transformers*, [Accessed 03-03-2023], 2022. [Online]. Available: https://www.youtube.com/watch?v=QvkQ1B3FBqA.

[120] L. Voita, *Transformer: Attention is All You Need*, https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html, [Accessed 27-09-2023], 2023.

[121] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—a contemplative retrospection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106 126, 2023.

[122] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[123] I. Goodfellow, "Nips 2016 tutorial: generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[124] J. Bromley, I. Guyon, Y. LeCun, E. Sackinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[125] D. Chicco, "Siamese neural networks: an overview," *Artificial neural networks*, pp. 73–94, 2021.

[126] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[127] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[128] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[129] G. Orrù, A. Piarulli, C. Conversano, and A. Gemignani, "Humanlike problemsolving abilities in large language models using chatgpt," *Frontiers in Artificial Intelligence*, vol. 6, p. 1 199 350, 2023.

[130] W. X. Zhao, K. Zhou, J. Li, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[131] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[132] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," *Nature Machine Intelligence*, vol. 2, no. 9, pp. 500–508, 2020.

[133] A. Mabrouk, R. P. Diaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, "Pneumonia detection on chest x-ray images using ensemble of deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 13, p. 6448, 2022.

[134] K. Ajitesh, *Ensemble Methods in Machine Learning: Examples - Data Analytics*, https://vitalflux.com/5-common-ensemble-methods-in-machine-learning/, [Accessed 09-08-2023], 2023.

[135] L. Derczynski, "Complementarity, f-score, and nlp evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 261–266.

[136] A. Chen, G. Stanovsky, S. Singh, and M. Gardner, "Evaluating question answering evaluation," in *Proceedings of the 2nd workshop on machine reading for question answering*, 2019, pp. 119–124.

[137] 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*, 2002, pp. 311–318.

[138] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[139] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013.

[140] J. Briggs, *The Ultimate Performance Metric in NLP*, https://towardsdatascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460, [Accessed 08-08-2023], 2021.

[141] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[142] A. Mohammadshahi, T. Scialom, M. Yazdani, *et al.*, "RQUGE: reference-free metric for evaluating question generation by answering the question," in *Findings of the Association for Computational Linguistics: ACL 2023*, Jul. 2023, pp. 6845–6867.

[143] X. Wang, C.-X. Yuan, B. Xu, and Z. Yu, "Diabetic foot ulcers: classification, risk factors and management," *World Journal of Diabetes*, vol. 13, no. 12, p. 1049, 2022.

[144] L. Chen, S. Sun, Y. Gao, and X. Ran, "Global mortality of diabetic foot ulcer: a systematic review and meta-analysis of observational studies," *Diabetes, Obesity and Metabolism*, vol. 25, no. 1, pp. 36–45, 2023.

[145] R. G. Frykberg, "Diabetic foot ulcers: pathogenesis and management," *American family physician*, vol. 66, no. 9, p. 1655, 2002.

[146] C. McCague, K. MacKay, C. Welsh, A. Constantinou, R. Jena, and M. Crispin-Ortuzar, "Position statement on clinical evaluation of imaging ai," *The Lancet Digital Health*, vol. 5, no. 7, e400–e402, 2023.

[147] M. Monteiro-Soares and J. V. Santos, "Diabetes foot-related complications," International Diabetes Federation, Tech. Rep., 2022.

[148] M. H. Yap, B. Cassidy, and C. Kendrick, *Diabetic foot ulcers grand challenge*. Springer, 2022.

[149] A. Galdran, G. Carneiro, and M. A. G. Ballester, "Convolutional nets versus vision transformers for diabetic foot ulcer classification," in *Diabetic Foot Ulcers Grand Challenge*, 2022, pp. 21–29.

[150] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[151] A. Kolesnikov, L. Beyer, X. Zhai, *et al.*, "Big transfer (bit): general visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 2020, pp. 491–507.

[152] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 2019, pp. 6105–6114.

[153] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, 2021, pp. 10 347–10 357.

[154] S. H. Haji and A. M. Abdulazeez, "Comparison of optimization techniques based on gradient descent algorithm: a review," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 4, pp. 2715–2743, 2021.

[155] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[156] L. Bloch, R. Brüngel, and C. M. Friedrich, "Boosting efficientnets ensemble performance via pseudo-labels and synthetic images by pix2pixhd for infection and ischaemia classification in diabetic foot ulcers," in *Diabetic Foot Ulcers Grand Challenge*, 2022, pp. 30–49.

[157] D.-H. Lee *et al.*, "Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 896.

[158] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks.," *Commun. Acm*, vol. 63, no. 11, pp. 139–144, 2020.

[159] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[160] M. Ahsan, S. Naz, R. Ahmad, H. Ehsan, and A. Sikandar, "A deep learning approach for diabetic foot ulcer classification and recognition," *Information*, vol. 14, no. 1, p. 36, 2023.

[161] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, and M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: dataset and techniques," *Computers in Biology and Medicine*, vol. 117, p. 103 616, 2020.

[162] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[163] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[164] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 4278–4284, 2017.

[165] F. Santos, E. Santos, L. H. Vogado, *et al.*, "Dfu-vgg, a novel and improved vgg-19 network for diabetic foot ulcer classification," in *2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP)*, vol. CFP2255E-ART, 2022, pp. 1–4.

[166] E. Santos, F. Santos, J. Dallyson, K. Aires, J. M. R. S. Tavares, and R. Veras, "Diabetic foot ulcers classification using a fine-tuned cnns ensemble," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, 2022, pp. 282–287.

[167] A. Khandakar, M. E. Chowdhury, M. B. I. Reaz, *et al.*, "A machine learning model for early detection of diabetic foot using thermogram images," *Computers in biology and medicine*, vol. 137, p. 104 838, 2021.

[168] A. Qayyum, A. Benzinou, M. Mazher, and F. Meriaudeau, "Efficient multi-model vision transformer based on feature fusion for classification of dfuc2021 challenge," in *Diabetic*

Foot Ulcers Grand Challenge, M. H. Yap, B. Cassidy, and C. Kendrick, Eds., 2022, pp. 62–75.

[169] S. Ahmed and H. Naveed, "Bias adjustable activation network for imbalanced data—diabetic foot ulcer challenge 2021," in *Diabetic Foot Ulcers Grand Challenge*, M. H. Yap, B. Cassidy, and C. Kendrick, Eds., 2022, pp. 50–61.

[170] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O'Shea, D. Gillespie, and N. D. Reeves, "Analysis towards classification of infection and ischaemia of diabetic foot ulcers," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4.

[171] M. Tan and Q. Le, "Efficientnetv2: smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 10 096–10 106.

[172] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[173] A.-K. Duong, H.-L. Nguyen, and T.-T. Truong, "Large margin cotangent loss for deep similarity learning," in *2022 International Conference on Advanced Computing and Analytics (ACOMPA)*, 2022, pp. 40–47.

[174] M. A. jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using k- nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013.

[175] D. Sculley, G. Holt, D. Golovin, *et al.*, "Hidden technical debt in machine learning systems," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15, 2015, pp. 2503–2511.

[176] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, 2019.

[177] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 1214–1223.

[178] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[179] P. Pathak, J. Zhang, and D. Samaras, "Local learning on transformers via feature reconstruction," *arXiv preprint arXiv:2212.14215*, 2022.

[180] L. Kwak and H. Bai, *The role of federated learning models in medical imaging*, 2023.

[181] N. Rieke, J. Hancox, W. Li, *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, p. 119, 2020.

[182] M. F. Sohan and A. Basalamah, "A systematic review on federated learning in medical image analysis," *IEEE Access*, vol. 11, pp. 28 628–28 644, 2023.

[183] A. Rahman, M. S. Hossain, G. Muhammad, *et al.*, "Federated learning-based ai approaches in smart healthcare: concepts, taxonomies, challenges and open issues," *Cluster computing*, vol. 26, no. 4, pp. 2271–2311, 2023.

[184] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[185] M. J. Sheller, B. Edwards, G. A. Reina, *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, p. 12 598, 2020.

[186] M. Joshi, A. Pal, and M. Sankarasubbu, "Federated learning for healthcare domain-pipeline, applications and challenges," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 4, pp. 1–36, 2022.

[187] S. Kim and S. Lee, "Self–supervised augmentation of quality data based on classification-reinforced gan," in *2023 17th International Conference on Ubiquitous Information Management and Communication*, 2023, pp. 1–7.

[188] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–9, 2020.

[189] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari, "Machine learning and artificial intelligence in research and healthcare," *Injury*, vol. 54, S69–S73, 2023.

[190] D. Saraswat, P. Bhattacharya, A. Verma, *et al.*, "Explainable ai for healthcare 5.0: opportunities and challenges," *IEEE Access*, vol. 10, pp. 84 486–84 517, 2022.

[191] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[192] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[193] H. Doğruel, M. Aydemir, and M. K. Balci, "Management of diabetic foot ulcers and the challenging points: an endocrine view," *World Journal of Diabetes*, vol. 13, no. 1, p. 27, 2022.

[194] J. J. van Netten, D. Clark, P. A. Lazzarini, M. Janda, and L. F. Reed, "The validity and reliability of remote diabetic foot ulcer assessment using mobile phone images," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[195] M. H. Yap, C. Kendrick, N. D. Reeves, M. Goyal, J. M. Pappachan, and B. Cassidy, "Development of diabetic foot ulcer datasets: an overview," in *Diabetic Foot Ulcers Grand Challenge*, M. H. Yap, B. Cassidy, and C. Kendrick, Eds., Cham: Springer International Publishing, 2022, pp. 1–18.

[196] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang, "Dfu_qutnet: diabetic foot ulcer classification using novel deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 21-22, pp. 15 655–15 677, 2020.

[197] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, and Y. Duan, "Robust application of new deep learning tools: an experimental study in medical imaging," *Multimedia Tools and Applications*, pp. 1–29, 2022.

[198] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, *et al.*, "Towards a better understanding of transfer learning for medical imaging: a case study," *Applied Sciences*, vol. 10, no. 13, p. 4523, 2020.

[199] I. Cohen, Y. Huang, J. Chen, *et al.*, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[200] P. Khosla, P. Teterwak, C. Wang, *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 18 661–18 673. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[201] M. D. Li, K. Chang, B. Bearce, *et al.*, "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging," *NPJ digital medicine*, vol. 3, no. 1, p. 48, 2020.

[202] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[203] M. AbdulRaheem, I. Oladipo, S. Ajagbe, B. Balogun, and N. Emma-Adamah, "Continuous eye disease severity evaluation system using siamese neural networks," *ParadigmPlus*, vol. 4, pp. 1–17, Mar. 2023.

[204] M. N. Akbar, X. Wang, D. Erdoğmuş, and S. Dalal, "Penet: continuous-valued pulmonary edema severity prediction on chest x-ray using siamese convolutional networks," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2022, pp. 1834–1838.

[205] J. Fiaidhi, S. Mohammed, and P. Zezos, "Siamese neural network for labeling severity of ulcerative colitis video colonoscopy: a thick data approach," in *Intelligent Systems*

and Applications: Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 1, Springer, 2022, pp. 124–135.

[206] Z. Hu, I. Dychka, Y. Sulema, Y. Valchuk, and O. Shkurat, "Method of medical images similarity estimation based on feature analysis," *International Journal of Intelligent Systems and Applications*, vol. 10, no. 5, pp. 14–22, 2018.

[207] M. Ionescu, A. D. Glodeanu, I. R. Marinescu, A. G. Ionescu, and C. C. Vere, "Similarity analysis for medical images using color and texture histogramss," *Current Health Sciences Journal*, vol. 48, no. 2, pp. 196–202, 2022.

[208] T. R. Ornob, G. Roy, and E. Hassan, "Covidexpert: a triplet siamese neural network framework for the detection of covid-19," *Informatics in Medicine Unlocked*, vol. 37, p. 101 156, Jan. 2023.

[209] A. Mehmood, M. Maqsood, M. Bashir, and Y. Shuyuan, "A deep siamese convolution neural network for multi-class classification of alzheimer disease," *Brain Sciences*, vol. 10, no. 2, p. 84, Feb. 2020. [Online]. Available: http://dx.doi.org/10.3390/brainsci10020084.

[210] X. Zeng, H. Chen, Y. Luo, and W. Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30 744–30 753, 2019.

[211] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[212] M. E. Vasconcellos, B. G. Ferreira, J. S. Leandro, *et al.*, "Siamese convolutional neural network for heartbeat classification using limited 12-lead ecg datasets," *Ieee Access*, vol. 11, pp. 5365–5376, 2023.

[213] J. Wang, Z. Fang, N. Lang, H. Yuan, M.-Y. Su, and P. Baldi, "A multi-resolution approach for spinal metastasis detection using deep siamese neural networks," *Computers in biology and medicine*, vol. 84, pp. 137–146, 2017.

[214] F. Hajamohideen, N. Shaffi, M. Mahmud, *et al.*, "Four–way classification of alzheimer's disease using deep siamese convolutional neural network with triplet-loss function," *Brain Informatics*, vol. 10, no. 1, pp. 1–13, 2023.

[215] S. Tummala and A. K. Suresh, "Few-shot learning using explainable siamese twin network for the automated classification of blood cells," *Medical & Biological Engineering & Computing*, pp. 1–15, 2023.

[216] M. Shorfuzzaman and M. S. Hossain, "Metacovid: a siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients," *Pattern recognition*, vol. 113, p. 107 700, 2021.

[217] S. Ahuja, B. K. Panigrahi, N. Dey, A. Taneja, and T. K. Gandhi, "Mcs-net: multi-class siamese network for severity of covid-19 infection classification from lung ct scan slices," *Applied Soft Computing*, vol. 131, p. 109 683, 2022.

[218] J. H. Cueva, D. Castillo, H. Espinós-Morató, D. Durán, P. Díaz, and V. Lakshminarayanan, "Detection and classification of knee osteoarthritis," *Diagnostics*, vol. 12, no. 10, p. 2362, 2022.

[219] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[220] L. Cai, S. Zhou, X. Yan, and R. Yuan, "A stacked bilstm neural network based on coattention mechanism for question answering," *Computational intelligence and neuroscience*, vol. 2019, 2019.

[221] D. Dayya, O. J. O'Neill, T. B. Huedo-Medina, N. Habib, J. Moore, and K. Iyer, "Debridement of diabetic foot ulcers," *Advances in Wound Care*, vol. 11, no. 12, pp. 666–686, 2022.

[222] C. Almonaci Hernández, K. Juarez-Moreno, M. Castañeda-Juarez, H. Almanza-Reyes, A. Pestryakov, N. Bogdanchikova, *et al.*, "Silver nanoparticles for the rapid healing of diabetic foot ulcers," *Int. J. Med. Nano Res*, vol. 4, no. 01910.23937, pp. 2378–3664, 2017.

[223] M. Tobalem and I. Uçkay, "Evolution of a diabetic foot infection," *New England Journal of Medicine*, vol. 369, no. 23, pp. 2252–2252, 2013.

[224] R. Kamalraj, S. Neelakandan, M. R. Kumar, V. C. S. Rao, R. Anand, and H. Singh, "Interpretable filter based convolutional neural network (if-cnn) for glucose prediction and classification using pd-ss algorithm," *Measurement*, vol. 183, p. 109 804, 2021.

[225] B. Ploderer, D. Clark, R. Brown, J. Harman, P. A. Lazzarini, and J. J. Van Netten, "Self-monitoring diabetes-related foot ulcers with the myfootcare app: a mixed methods study," *Sensors*, vol. 23, no. 5, p. 2547, 2023.

[226] A. Esteva, B. Kuprel, R. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, Jan. 2017.

[227] R. Mehr and A. Ameri, "Skin cancer detection based on deep learning," *Journal of biomedical physics and engineering*, vol. 12, pp. 559–568, Dec. 2022.

[228] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[229] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[230] B. Meskó, G. Hetényi, and Z. Győrffy, "Will artificial intelligence solve the human resource crisis in healthcare?" *BMC health services research*, vol. 18, no. 1, pp. 1–4, 2018.

[231] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[232] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[233] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[234] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[235] L. He, A. Bian, and M. Jaggi, "Cola: decentralized linear learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[236] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[237] I. Hegedűs, G. Danner, and M. Jelasity, "Decentralized learning works: An empirical comparison of gossip learning and federated learning," *Journal of Parallel and Distributed Computing*, vol. 148, pp. 109–124, 2021.

[238] D. Mäenpää, *Towards peer–to–peer federated learning: algorithms and comparisons to centralized federated learning*, 2021.

[239] Y. Zhang, Y. Lv, and F. Liu, "A systematic survey for differential privacy techniques in federated learning," *Journal of Information Security*, vol. 14, no. 2, pp. 111–135, 2023.

[240] J. Konevcny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[241] J. Qi, Q. Zhou, L. Lei, and K. Zheng, "Federated reinforcement learning: techniques, applications, and open challenges," *Intelligence & Robotics*, vol. 1, Oct. 2021.

[242] D. Sirohi, N. Kumar, P. S. Rana, S. Tanwar, R. Iqbal, and M. Hijjii, "Federated learning for 6g-enabled secure communication systems: a comprehensive survey," *Artificial Intelligence Review*, pp. 1–93, 2023.

[243] P. Vanhaesebrouck, A. Bellet, and M. Tommasi, "Decentralized collaborative learning of personalized models over networks," in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 509–517.

[244] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Reviewing federated machine learning and its use in diseases prediction," *Sensors*, vol. 23, no. 4, p. 2112, 2023.

[245] R. Kontar, N. Shi, X. Yue, *et al.*, "The internet of federated things," *IEEE Access*, vol. 9, pp. 156 071–156 113, 2021.

[246] R. Shokri and V. Shmatikov, "Privacy–preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[247] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4289–4301, 2022.

[248] M. Asad, A. Moustafa, T. Ito, and M. Aslam, "Evaluating the communication efficiency in federated learning algorithms," in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2021, pp. 552–557.

[249] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi, "Don't use large mini-batches, use local sgd," *arXiv preprint arXiv:1808.07217*, 2018.

[250] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "Braintorrent: a peer-to-peer environment for decentralized federated learning," *arXiv preprint arXiv:1905.06731*, 2019.

[251] M. N. Fekri, K. Grolinger, and S. Mir, "Distributed load forecasting using smart meter data: federated learning with recurrent neural networks," *International Journal of Electrical Power and Energy Systems*, vol. 137, p. 107 669, 2022.

[252] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[253] C. Alampalle, S. Hegde, S. Jahagirdar, and S. Gangisetty, "Weakly supervised visual question answer generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5588–5596.

[254] S. Gottlieb and L. Silvis, "How to Safely Integrate Large Language Models Into Health Care," *JAMA Health Forum*, vol. 4, no. 9, e233909–e233909, Sep. 2023.

[255] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," *Foundations and Trends® in Information Retrieval*, vol. 13, no. 2–3, pp. 127–298, 2019. [Online]. Available: http://dx.doi.org/10.1561/1500000074.

[256] A. Dhandapani and V. Vadivel, "Template–Based Question Answering System Over the Semantic Web," *International Journal of Information Retrieval Research*, vol. 12, pp. 1–17, Sep. 2022.

[257] S. Singh and S. Susan, "Healthcare Question–Answering System: Trends and Perspectives," in May 2023, pp. 239–249.

[258] Haque, Ahshanul and Chowdhury, Md Naseef–Ur–Rahman and Soliman, Hamdy, "Transforming chronic disease management with chatbots: key use cases for personalized and cost-effective care," in *2023 Sixth International Symposium on Computer, Consumer and Control (IS3C)*, IEEE, 2023, pp. 367–370.

[259] S. Montagna, S. Ferretti, L. C. Klopfenstein, A. Florio, and M. F. Pengo, "Data decentralisation of llm-based chatbot systems in chronic disease self–management," in *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, 2023, pp. 205–212.

[260] G. K. Vamsi, A. Rasool, and G. Hajela, "Chatbot: A deep neural network based human to machine conversation model," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, pp. 1–7.

[261] R. Kaladevi, S. Saidineesha, P. K. Priya, K. Nithiya, and S. S. Gayatri, "Chatbot for healthcare using machine learning," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, 2023, pp. 1–4.

[262] S. Dipanjan, *A practitioner's guide to nlp*, https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72, [Accessed 31-08-2023], 2018.

[263] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information fusion*, vol. 36, pp. 10–25, 2017.

[264] T. A. Team, *Nlp concepts and workflow*, https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7, [Accessed 31-08-2023], 2020.

[265] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[266] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997.

[267] 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*, 2014, pp. 1532–1543.

[268] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[269] Chan, Ying-Hong and Fan, Yao-Chung, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd workshop on machine reading for question answering*, 2019, pp. 154–162.

[270] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li, "Neural generative question answering," *arXiv preprint arXiv:1512.01337*, 2015.

[271] Y. Zhang and Z. Xu, "Bert for question answering on squad 2.0," *Stanford University Report*, 2019.

[272] X. Zhang, J. Wu, Z. He, X. Liu, and Y. Su, "Medical exam question answering with large-scale reading comprehension," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[273] H. Zhou, B. Lei, Z. Liu, and Z. Liu, "Dut-BIM at MEDIQA 2019: utilizing transformer network and medical domain-specific contextualized representations for question answering," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 446–452. [Online]. Available: https://aclanthology.org/W19-5047.

[274] Q. Bao, L. Ni, and J. Liu, "Hhh: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention," Feb. 2020, pp. 1–10.

[275] N. Harilal, R. Shah, S. Sharma, and V. Bhutani, "Caro: an empathetic health conversational chatbot for people with major depression," Jan. 2020, pp. 349–350.

[276] S. Soni and K. Roberts, "Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 5532–5538. [Online]. Available: https://aclanthology.org/2020.lrec-1.679.

[277] J. Alzubi, R. Jain, A. Singh, P. Parwekar, and M. Gupta, "Cobert: covid-19 question answering system using bert," *Arabian Journal for Science and Engineering*, Jun. 2021.

[278] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Question-aware transformer models for consumer health question summarization," *Journal of Biomedical Informatics*, vol. 128, p. 104 040, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046422000569.

[279] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extractio: issues and methods," *Natural Language Engineering*, vol. 26, no. 3, pp. 259–291, 2020.

[280] G. Maarten, *Keybert: minimal keyword extraction with bert.* 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265.

[281] M. Kulkarni, D. Mahata, R. Arora, and R. Bhowmik, "Learning rich representation of keyphrases from text," *arXiv preprint arXiv:2112.08547*, 2021.

[282] M. Q. Khan, A. Shahid, M. I. Uddin, *et al.*, "Impact analysis of keyword extraction using contextual word embedding," *PeerJ Computer Science*, vol. 8, e967, 2022.

[283] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," Jan. 2018, pp. 3901–3910.

[284] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng, "A review on question generation from natural language text," *ACM Transactions on Information Systems*, vol. 40, pp. 1–43, Jan. 2022.

[285] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[286] I. Beltagy, K. Lo, and A. Cohan, "Scibert a pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620.

[287] A. Costello, E. Fedorova, Z. Jin, and R. Mihalcea, "Editing a woman's voice," *arXiv preprint arXiv:2212.02581*, 2022.

[288] Y. Gu, R. Tinn, H. Cheng, *et al.*, *Domain-specific language model pretraining for biomedical natural language processing*, 2020. eprint: arXiv:2007.15779.

[289] J. Lee, W. Yoon, S. Kim, *et al.*, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics (Oxford, England)*, vol. 36, Sep. 2019.

[290] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[291] M. Yasunaga, J. Leskovec, and P. Liang, "Linkbert: pretraining language models with document links," *arXiv preprint arXiv:2203.15827*, 2022.

[292] ParseHub. "The most powerful web scraper." (2022), [Online]. Available: https%5C://www.parsehub.com/ (visited on 01/30/2022).

[293] Octoparse. "Easy web scraping for anyone." (2019), [Online]. Available: https%5C://www.octoparse.com/ (visited on 01/30/2022).

[294] W. Qi, Y. Yan, Y. Gong, *et al.*, "Prophetnet: predicting future n-gram for sequence-to-sequence pre-training," *arXiv preprint arXiv:2001.040632*, 2020.

[295] A. Otegi, J. A. Campos, G. Azkune, A. Soroa, and E. Agirre, "Automatic evaluation vs user preference in neural textual questionanswering over covid–19 scientific literature," in *Workshop on NLP for COVID–19 (Part 2) at EMNLP 2020*, 2020.

[296] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 784–789. [Online]. Available: https://aclanthology.org/P18-2124.

[297] E. Choi, H. He, M. Iyyer, *et al.*, "Quac– question answering in context," *arXiv preprint arXiv:1808.07036*, 2018.

[298] J. Pereira, R. Fidalgo, R. Lotufo, and R. Nogueira, "Visconde: multi-document qa with gpt-3 and neural reranking," in *European Conference on Information Retrieval*, Springer, 2023, pp. 534–543.

[299] A. Sharma and S. Kumar, "Machine learning and ontology-based novel semantic document indexing for information retrieval," *Computers and Industrial Engineering*, vol. 176, p. 108 940, 2023.

[300] A. Sauer, R. B. Gramacy, and D. Higdon, "Active learning for deep gaussian process surrogates," *Technometrics*, vol. 65, no. 1, pp. 4–18, 2023.

[301] A. N. Klonoff, W.-.-A. Lee, N. Y. Xu, K. T. Nguyen, A. DuBord, and D. Kerr, "Six digital health technologies that will transform diabetes," *Journal of diabetes science and technology*, vol. 17, no. 1, pp. 239–249, 2023.

[302] R. Shrestha, K. Kafle, and C. Kanan, "An investigation of critical issues in bias mitigation techniques," in *Proceedings of the IEEE–CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1943–1954.

[303] R. Wang, P. Chaudhari, and C. Davatzikos, "Bias in machine learning models can be significantly mitigated by careful training: evidence from neuroimaging studies," *Proceedings of the National Academy of Sciences*, vol. 120, no. 6, e2211613120, 2023.

[304] T. Hulsen, "Explainable Artificial Intelligence XAI: Concepts and Challenges in Healthcare," *AI*, vol. 4, no. 3, pp. 652–666, 2023.

[305] K. T. Karason, D. Vo, J. Grauslund, and M. L. Rasmussen, "Comparison of different methods of retinal imaging for the screening of diabetic retinopathy: a systematic review," *Acta Ophthalmologica*, vol. 100, no. 2, pp. 127–135, 2022.

[306] T. L. D. Endocrinology, "Diabetes care and ai– a looming threat or a necessary advancement?" *The Lancet Diabetes Endocrinology*, vol. 11, no. 7, p. 441, 2023.

[307] S. R. Joshua, S. Shin, J.-H. Lee, and S. K. Kim, "Health to eat: a smart plate with food recognition, classification, and weight measurement for type–2 diabetic mellitus patients' nutrition control," *Sensors*, vol. 23, no. 3, p. 1656, 2023.

[308] G. Latif, N. Mohammad, and J. Alghazo, "Deepfruit: a dataset of fruit images for fruit classification and calories calculation," *Data in Brief*, p. 109 524, 2023.

[309] G. Fagherazzi, "Technologies will not make diabetes disappear: how to integrate the concept of diabetes distress into care," *Diabetes Epidemiology and Management*, p. 100 140, 2023.

[310] WHO, "Global report on diabetes," World Health Organisation, 2016. [Online]. Available: https://www.who.int/publications/i/item/9789241565257.

Un Ecosystème Innovant basé sur Deep Learning: Contributions pour la Prévention et la Prédiction des Complications du Diabète

Résumé : En 2021, les estimations indiquaient qu'environ 537 millions de personnes étaient touchées par le diabète, un chiffre qui devrait grimper à 643 millions d'ici 2030, et encore à 783 millions d'ici 2045. Caractérisé comme une maladie métabolique persistante, le diabète nécessite des soins et une gestion quotidiens continus. Le fardeau des maladies chroniques pèse lourdement sur les systèmes de santé lorsqu'il affecte une partie substantielle de la population. De telles circonstances ont un impact négatif non seulement sur le bien-être général d'une grande partie de la population, mais contribuent également de manière significative aux dépenses de santé. Dans le contexte de Maurice, selon le rapport le plus récent de la Fédération Internationale du Diabète, la prévalence du diabète, en particulier du diabète de type 2 (T2D), était de 22,6 % de la population en 2021, avec des projections indiquant une hausse à 26,6 % d'ici 2045. Face à cette tendance alarmante, une évolution concomitante a été observée dans le domaine de la technologie, les techniques d'intelligence artificielle démontrant des capacités prometteuses dans les domaines de la médecine et de la santé. Cette thèse de doctorat entreprend l'exploration de l'intersection entre l'intelligence artificielle et l'éducation, la prévention, et la gestion du diabète.

Nous nous sommes d'abord concentrés sur l'exploration du potentiel de l'Intelligence Artificielle (IA) pour répondre à une complication critique liée au diabète - l'Ulcère du Pied Diabétique (DFU). L'émergence des DFU présente un risque grave d'amputations des membres inférieurs, entraînant des répercussions socio-économiques sévères. En réponse, nous avons proposé une solution innovante de classification, DFU-SIAM et DFU-HELPER. DFU-HELPER sert de mesure préliminaire pour valider les protocoles de traitement administrés par les professionnels de la santé aux patients individuels affectés par les DFU. L'évaluation initiale de l'outil proposé a montré des caractéristiques de performance prometteuses, bien que des affinements et des tests rigoureux soient impératifs. Les efforts collaboratifs avec les experts en santé publique seront essentiels pour évaluer l'efficacité pratique de l'outil dans des scénarios réels.

Notre recherche a également abordé les aspects critiques de la vie privée et de la confidentialité inhérents à la manipulation des données liées à la santé. Reconnaissant l'importance extrême de la protection des informations sensibles, nous nous sommes plongés dans le domaine de l'apprentissage fédéré Peer-to-Peer. Cette investigation a trouvé spécifiquement son application dans notre proposition pour l'outil DFU-SIAM présenté plus tôt. En explorant cette approche avancée, nous avons cherché à assurer que la mise en œuvre de notre technologie soit conforme aux normes de confidentialité, favorisant ainsi un environnement de confiance et sécurisé pour la gestion des données de santé

Enfin, recherche s'est étendue au développement d'un agent conversationnel intelligent conçu pour offrir un support 24 heures sur 24 aux personnes recherchant des informations sur le diabète. Dans la poursuite de cet objectif, la création d'un jeu de données approprié était primordiale. Dans ce contexte, nous avons utilisé des techniques de traitement du langage naturel pour sélectionner des données à partir de sources médias en ligne axées sur le contenu lié au diabète.

Mots clés : Intelligence Artificielle, Diabète, Ulcère du Pied Diabétique, Agent Conversationel, Apprentissage Profond, Réseau Neuronal Siamois, Apprentissage Fédéré, Confidentialité des Données, Prévention du Diabète, Prédictions des Complications du Diabète, Technologie et Santé.

An Innovative Ecosystem Based on Deep Learning: Contributions for the Prevention and Prediction of Diabetes Complications

Abstract: In the year 2021, estimations indicated that approximately 537 million individuals were affected by diabetes, a number anticipated to escalate to 643 million by the year 2030 and further to 783 million by 2045. Diabetes, characterized as a persistent metabolic ailment, necessitates unceasing daily care and management. The burden of chronic conditions is cumbersome on healthcare systems when it afflicts a substantial segment of the population. Such circumstances not only detrimentally impact the overall well-being of a significant population but also contribute significantly to healthcare expenditure. In the context of Mauritius, as per the most recent report by the International Diabetes Federation, the prevalence of diabetes, specifically Type 2 Diabetes (T2D), stood at 22.6% of the population in 2021, with projections indicating a surge to 26.6% by the year 2045. Amidst this alarming trend, a concurrent advancement has been observed in the realm of technology, with artificial intelligence techniques showcasing promising capabilities in the spheres of medicine and healthcare. This doctoral dissertation embarks on the exploration of the intersection between artificial intelligence and diabetes education, prevention, and management. We initially focused on exploring the potential of deep learning to address a critical complication linked to diabetes: Diabetic Foot Ulcer (DFU). The emergence of DFU poses the grave risk of lower limb amputations, consequently leading to severe socio-economic repercussions. In response, we put forth an innovative solution, DFU-SIAM for DFU classification and DFU-HLEPER. DFU-HELPER serves as a preliminary measure for validating the treatment protocols administered by healthcare professionals to individual patients afflicted by DFU. The initial assessment of the proposed tool has exhibited promising performance characteristics, although further refinement and rigorous testing are imperative.

Our research also addressed the critical aspects of privacy and confidentiality inherent in handling health-related data. Acknowledging the extreme importance of safeguarding sensitive information, we delved into the realm of Peer-to-Peer Federated Learning. This investigation specifically found application in our proposal for the DFU-SIAM discussed earlier.

Finally, our research extended to the development of an intelligent conversational agent designed to offer round-the-clock support for individuals seeking information about diabetes. In pursuit of this goal, the creation of an appropriate dataset was paramount. In this context, we leveraged Natural Language Processing techniques to curate data from online media sources focusing on diabetes-related content. Although further exploration is needed for the different experiments undertaken within this thesis, our efforts yielded compelling outcomes and insights regarding the integration of Artificial Intelligence (AI) to cultivate an AI-powered ecosystem dedicated to diabetes education, prevention, and management. Furthermore, we successfully highlighted the obstacles confronting researchers, notably the constraints arising from data availability and processing capabilities. An especially promising avenue for future research emerged—the systematic examination of model explainability. This aspect is of significant importance as it directly influences the acceptance of AI solutions by healthcare professionals.

Keywords: Diabetes, Artificial Intelligence, Diabetic Foot Ulcer, AI Chatbot, Deep Learning, Siamese Neural Network, Federated Learning, Data Privacy, Diabetes Prevention, Diabetes Complication Predictions, Technology and Healthcare

Univ. de Limoges, CNRS, XLIM, UMR 7252, F-87000 Limoges, France