

Development of the Virtual Doctor: A Walk-In Cabin for Automated Medical Data Acquisition and AI-Driven Decision Support

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf
vorgelegt von

Jan Benedikt Ruhland
aus München

Düsseldorf, Dezember, 2025

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Examiners:

1. Prof. Dr. Dominik Heider
2. Prof. Dr. Gunnar W. Klau

Tag der mündlichen Prüfung: **28.05.2026**

Acknowledgements

I would like to express my sincere gratitude to everyone who supported me throughout the journey of my dissertation.

First and foremost, I extend my deepest thanks to my supervisor, Prof. Dr. Dominik Heider, for his exceptional guidance, support, and patience. His expertise and thoughtful advice were invaluable in shaping my research and helping me navigate both the technical and conceptual challenges of this work. I am equally grateful for the moral support and encouragement he provided throughout the process, which helped me stay motivated and confident during difficult moments. I am truly thankful for the opportunity to have learned from him.

I would also like to extend my gratitude to Prof. Dr. Gunnar W. Klau for serving as my mentor and for taking the time to review my thesis. I appreciate his willingness to support this work in his role as mentor and his commitment to ensuring a thorough and careful evaluation of my thesis. His involvement in the review process contributed meaningfully to the completion of this project, and I am grateful for his time and consideration.

I am also thankful for the collaborative spirit and encouragement of my colleagues and friends within our research group. In particular, I would like to thank Leon Fehse, Jan Henric Klau, and Dr. Roman Martin for their helpful feedback, critical insights, and contributions to the projects presented in this dissertation. Their support and constructive discussions played a significant role in improving the quality of my work.

My heartfelt thanks go to my wife Laura, whose constant support, understanding, and belief in me have been a cornerstone of strength throughout this journey. A special mention also goes to our son Benedikt, who joined us near the end of this process and brought with him joy and motivation that fueled my final efforts.

I am deeply grateful to my mother Hertha for her unwavering support and encouragement throughout my academic path and beyond. Her faith in my abilities has been a constant source of inspiration and determination.

I would also like to acknowledge my friends and extended family for their invaluable support, whether through thoughtful conversations, emotional encouragement, or simply being there when I needed a moment of rest. In particular, I would like to thank Dr. Wolfgang Büsing, and Gretel and Reinhard Arndt for their kindness and steady presence during this time.

Lastly, I am thankful to the broader academic community, the researchers and practitioners whose work has inspired and informed my own. Their contributions laid the foundation on which this dissertation was built.

Eidesstattliche Erklärung/Declaration

Ich, *Jan Benedikt Ruhland*, versichere an Eides Statt, dass die Dissertation „*Development of the Virtual Doctor: A Walk-In Cabin for Automated Medical Data Acquisition and AI-Driven Decision Support*“ von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Die Dissertation wurde in der vorgelegten oder einer ähnlichen Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 26.11.2025

Jan Benedikt Ruhland

Abstract

The German healthcare system is increasingly under pressure due to demographic changes and significant disparities in medical care between urban and rural regions. The COVID-19 pandemic in 2019 highlighted the structural weaknesses of the system and further exacerbated the burden on healthcare personnel.

A promising approach to alleviate the workload of healthcare staff, reduce regional inequalities in medical care, and promote personalized diagnostics is the targeted use of Artificial Intelligence (AI) in medicine. Although AI is already established in certain applications, its full potential in everyday clinical practice remains largely untapped.

This dissertation presents four projects that focus on the development and integration of AI-based methods into the concept of the “Virtual Doctor” cabin. This cabin is equipped with various biomedical sensors and uses AI models to enable automated health assessments of patients. The design, technical implementation, and integration of hardware and software components for the realization of the “Virtual Doctor” cabin constitute the overarching project, into which the three other projects are embedded.

The second project addresses the methodological foundations of deep learning. A particular focus is on the application of adaptive normalization techniques in neural networks, which are systematically evaluated using diverse medical datasets. This methodology improves the adaptation of models to heterogeneous clinical data, thereby enhancing the accuracy of diagnostic procedures.

The third project focuses on the development of an AI system for the automated detection of pathological features in retinal fundus images. The images were collected as part of a clinical study and enriched with annotation data to enable robust training and validation of the model. The findings from this project were directly integrated into the Virtual Doctor system and can be used in the future to support clinical decision-making in ophthalmology.

The fourth project deals with the implementation of a speech-based anamnesis system. A chatbot was developed to conduct structured anamnesis interviews with patients and to provide the collected information in a standardized format for healthcare professionals. Additionally, a virtual avatar based on generative image models was integrated to make interactions with the system more natural and intuitive.

Zusammenfassung

Das deutsche Gesundheitssystem steht zunehmend aufgrund demografischer Veränderungen sowie deutlicher Unterschiede in der medizinischen Versorgung zwischen städtischen und ländlichen Regionen unter Druck. Die COVID-19-Pandemie im Jahr 2019 hat die strukturellen Schwächen des Systems offen sichtbar gemacht und die Überlastung des medizinischen Personals weiter verstärkt.

Ein vielversprechender Ansatz zur Entlastung des medizinischen Personals, zur Reduktion regionaler Ungleichheiten in der Gesundheitsversorgung und zur Förderung einer personalisierten Diagnostik ist der gezielte Einsatz von Künstlicher Intelligenz (KI) in der Medizin. Obwohl KI bereits in bestimmten Anwendungsfeldern etabliert ist, bleibt das volle Potenzial dieser Technologie im klinischen Alltag bislang weitgehend ungenutzt.

Im Rahmen dieser Dissertation werden vier Projekte vorgestellt, die die Entwicklung und Integration KI-gestützter Verfahren in das Konzept der „Virtual Doctor“-Kabine zum Ziel haben. Diese Kabine ist mit verschiedenen biomedizinischen Sensoren ausgestattet und nutzt KI-Modelle, um eine automatisierte Gesundheitsbewertung von Patientinnen und Patienten zu ermöglichen. Die Konzeption, technische Umsetzung und Integration von Hardware- und Softwarekomponenten zur Realisierung der „Virtual Doctor“-Kabine bildet das übergeordnete Projekt, in das die drei weiteren Projekte eingebettet sind.

Das zweite Projekt behandelt die methodischen Grundlagen des Deep Learning. Ein Schwerpunkt liegt auf der Anwendung eines adaptiven Normalisierungsverfahrens in neuronalen Netzwerken, das anhand unterschiedlicher medizinischer Datensätze systematisch untersucht wird. Diese Methodik trägt zu einer verbesserten Anpassung der Modelle an heterogene klinische Daten bei und erhöht dadurch die Genauigkeit diagnostischer Verfahren.

Das dritte Projekt widmet sich der Entwicklung eines KI-Systems zur automatisierten Erkennung pathologischer Merkmale in retinalen Fundusaufnahmen. Die Bilder wurden im Rahmen einer klinischen Studie erhoben und durch Annotationsdaten ergänzt, wodurch ein fundierter Trainings- und Validierungsprozess ermöglicht wird. Die gewonnenen Erkenntnisse flossen direkt in das „Virtual Doctor“-System ein und können dort künftig für klinische Entscheidungshilfen in der augenärztlichen Diagnostik genutzt werden.

Das vierte Projekt beschäftigt sich mit der Implementierung eines sprachbasierten Anamnesesystems. Zu diesem Zweck wurde ein Chatbot entwickelt, der strukturierte anamnestische Gespräche mit Patientinnen und Patienten führt und die erfassten Informationen für medizinisches Fachpersonal standardisiert aufbereitet. Ergänzend wurde ein virtueller Avatar auf Grundlage generativer Bildmodelle integriert, der die Interaktion mit dem System menschlicher und intuitiver gestaltet.

Contents

1. Introduction	1
2. Background	4
2.1. Medical Examinations	4
2.1.1. Fundus Images	4
2.1.2. Diagnostics using Fundus Images	5
2.1.3. Skin Lesion Examinations	7
2.1.4. Type II Diabetes Mellitus	9
2.1.5. Anamnesis and Measures of Well-Being	10
2.1.5.1. Pulse Oximetry	11
2.1.5.2. Thermal Camera	13
2.1.6. Medical Studies	14
2.2. Machine Learning	14
2.2.1. Camera Calibration	15
2.2.1.1. Radial Distortion	16
2.2.1.2. Tangential Distortion	17
2.2.2. Deep Learning	17
2.2.2.1. Gradient Descent and Optimization Techniques	19
2.2.2.2. Multilayer Perceptron and Activation Functions	20
2.2.2.3. Convolutional Neural Networks	21
2.2.2.4. Normalization Approaches	23
2.2.2.5. Transformer	24
2.2.2.6. Data Augmentation	25
2.2.2.7. Evaluation Metrics	26
2.2.2.8. Model Calibration	28
3. Materials and Methods	30
3.1. Concept of the Virtual Doctor Diagnostic Cabin	31
3.1.1. Hardware Design	31
3.1.2. Distributed Systems Architecture	33
3.1.3. Database Structure	34
3.1.4. Software Interface	35
3.2. Decision Support using Deep Learning	36
3.2.1. Normalization Strategies	37

3.2.2. Pathological Anomalies in Fundus Images	38
3.3. The Virtual Doctor Avatar	38
3.4. Ethical Approvals	39
4. Results	41
4.1. Virtual Doctor Hardware and Software Framework	41
4.1.1. Contributions	42
4.1.2. Introduction	43
4.1.3. Materials and methods	44
4.1.4. Results	48
4.1.5. Discussion	49
4.1.6. Conclusion	51
4.1.7. Appendix	52
4.2. Adaptive Normalization Layer	60
4.2.1. Contributions	60
4.2.2. Introduction	61
4.2.3. Materials and methods	62
4.2.4. Results	65
4.2.5. Discussion	65
4.2.6. Conclusion	67
4.3. Pathological Anomaly Detection in Fundus Images	70
4.3.1. Contributions	70
4.3.2. Introduction	72
4.3.3. Materials and methods	76
4.3.4. Results	82
4.3.5. Discussion	87
4.3.6. Conclusion	88
4.3.7. Appendix	89
4.4. Development of the Virtual Doctor Avatar and Chatbot	97
4.4.1. Contributions	97
4.4.2. Introduction	99
4.4.3. Materials and methods	103
4.4.4. Results	109
4.4.5. Discussion	112
4.4.6. Conclusion	114
4.4.7. Appendix	115
5. Discussion	130
6. Conclusion	134

A. Challenges in Health Care	147
A.1. Demographic Shifts	147
A.2. Medical Implication of an Aging Population	148
A.3. Temporal Variation in Disease Incidence	149
B. Machine Learning Examples	151
B.1. Camera Calibration	151
C. The Virtual Doctor Hardware and Software	155
C.1. Software Modules and 3D Design	155
C.2. Circuit Design	156
C.3. Materials and Cost	157
C.4. Explainable AI	157
D. Curriculum Vitae	160
E. Data Availability	161
F. Ethical Approval	162

List of Figures

2.1. Biological Structure of a Fundus Image	5
2.2. Fundus Images with Pathological Regions	7
2.3. Skin Lesion Examination	8
2.4. Pulse Oximetry	12
2.5. Biologically inspired neural Network	18
2.6. Example of a corner detection filter	22
2.7. Image Augmentation Techniques	26
3.1. Biomedical Recordings within the Virtual Doctor	32
3.2. Database ER diagram	35
A.1. Demographic Trends	147
A.2. Healthy Life Expectancy	149
A.3. Melanoma Incidences	150
B.1. Lens System	151
C.1. Virtual Doctor Cabin Design	156
C.2. PCB Designs	157
C.3. Explainability of the T2DM model	159

List of Tables

2.1. T2DM Risk Markers	10
C.1. Software Modules	155
C.2. Materials and Costs	158

List of Abbreviations

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AMD	Age-related Macular Degeneration
API	Application Programming Interface
BMI	Body Mass Index
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
DR	Diabetic Retinopathy
DSGVO	Datenschutz-Grundverordnung
ER	Entity-Relationship
FN	False Negative
FP	False Positive
GELU	Gaussian Error Linear Unit
KI	Künstliche Intelligenz
LLM	Large Language Model
MCC	Matthews Correlation Coefficient
MLP	Multilayer Perceptron
NLP	Natural Language Processing
PCB	Printed Circuit Board
PDF	Probability Density Function
ReLU	Rectified Linear Unit
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
rPPG	Remote Photoplethysmography
SGD	Stochastic Gradient Descent
SpO ₂	Peripheral Capillary Oxygen Saturation
SVD	Singular Value Decomposition
T2DM	Type 2 Diabetes Mellitus
TN	True Negative
TP	True Positive
WHO	World Health Organization
WPA2	Wifi Protected Access 2

1. Introduction

Healthcare systems worldwide are under increasing pressure due to significant demographic shifts, most notably the aging of the global population [1]. As life expectancy increases, so does the prevalence of age-related and chronic illnesses, including cardiovascular disease, type 2 diabetes, and neurodegenerative disorders [2]. These conditions require continuous and often complex medical care, leading to a surge in demand for long-term treatment, hospital-based services, and individualized healthcare strategies. This strain is especially pronounced in rural and underserved areas, where healthcare facilities are often sparse, medical professionals are in short supply, and access to specialized diagnostic and therapeutic services is limited. This inequity became particularly evident during the Coronavirus Disease 2019 (COVID-19) pandemic, where delayed access to critical care and limited public health infrastructure in rural regions exacerbated outcomes and highlighted existing systemic vulnerabilities [3].

While the demographic shift presents an urgent and well-recognized challenge, formulating effective countermeasures is not as straightforward as it may initially appear. One intuitive solution is to increase the number of medical personnel, thereby addressing immediate shortages and supporting the care needs of an aging population. However, a deeper analysis of demographic structures reveals a more complex issue.

Data from the World Health Organization (WHO) [4] and Germany’s Federal Statistical Office [5] indicate that much of the demographic shift is driven by the population aged between 50 and 65 years. This demographic bulge has created a temporary but substantial increase in the elderly population with several implications regarding the labor market and increased pressure on healthcare and social expenditure systems [6]. Expanding the medical workforce may effectively manage this short-term surge in demand. However, once this generation passes away, the healthcare system could be left with an oversupply of medical staff.

Such overstaffing would not only create inefficiencies but could also pose a significant economic burden on national healthcare systems. This structural dilemma highlights the need for adaptive, forward-looking healthcare strategies rather than static workforce expansion. A qualitative assessment of this challenge is provided in A.1 and A.2.

In addition to the demographic challenges facing modern healthcare systems, there is a notable rise in the incidence of acute and environmentally influenced health conditions. Among these, skin cancer [7] and heat-related illnesses such as heat stroke are of particular

concern [8]. These conditions are increasingly being linked to climate change, which contributes to higher ambient temperatures, increased air pollution, and more frequent and intense heat waves [8, 9, 10, 11, 12, 13].

Statistical trends indicating the growing prevalence of these conditions, along with evidence correlating their incidence to environmental and climatic factors, are detailed in A.3. These emerging health risks place additional pressure on already overextended healthcare systems. They underscore the need for integrated preventive strategies, improved access to early diagnostic technologies, and public health interventions aimed at climate resilience and health equity.

The COVID-19 pandemic starkly exposed the vulnerabilities of existing healthcare infrastructures [3, 14]. The crisis revealed systemic weaknesses in staffing, resource allocation, and emergency preparedness. It triggered widespread calls not only for improved working conditions within the healthcare sector but also for the development of robust contingency plans for future global health emergencies [14, 15]. Although COVID-19 was a unique event, the broader structural challenges it exposed remain relevant. When extrapolated into the future, current trends suggest a worsening of the healthcare burden unless comprehensive reforms are undertaken.

However, defining and implementing effective countermeasures remains a complex undertaking. The multifaceted nature of the problem, spanning demographic shifts, chronic disease management, healthcare workforce planning, and climate-induced health risks, requires adaptive, context-specific strategies. Simple, one-dimensional solutions are unlikely to be sustainable. Instead, evidence-based approaches that integrate technological innovation, policy reform, and equitable access to care will be essential to building resilient and future-ready healthcare systems.

In response to the growing challenges facing global healthcare systems, technological transformation, particularly the integration of AI, is emerging as a promising strategy to enhance both the efficiency and quality of healthcare delivery [16]. AI technologies offer the potential to support clinical decision-making, automate repetitive administrative and diagnostic tasks, enable early disease detection through pattern recognition, and facilitate personalized treatment planning based on individual patient data.

Despite this significant potential, the real-world deployment of AI in clinical settings is still in its early stages. One of the primary barriers is the requirement for large, high-quality datasets to train and validate AI models, data that is often difficult to access due to stringent privacy and data protection regulations [17, 18]. Furthermore, the effectiveness of AI applications in healthcare is closely tied to ongoing technological advancements in areas such as sensor hardware, data acquisition, and computational infrastructure [19].

This thesis introduces a novel approach to AI-based decision support through the development of a semi-autonomous data acquisition system. The project culminated in the

creation of the Virtual Doctor diagnostic walk-in cabin, which integrates both hardware and software components to facilitate remote, AI-supported medical assessments. Throughout the development process, several advancements in deep learning were implemented, including adaptive normalization techniques and functional localization mechanisms, which significantly improved the performance and interpretability of the AI models used for diagnostic decision support.

The final phase of the project focused on transforming the traditional anamnesis process using a Large Language Model (LLM). This transition enabled fully automated, personalized patient interviews, where patient responses are processed, interpreted, and summarized into structured anamnesis reports. These reports are designed to assist the attending physician by providing clinically relevant insights, thereby improving diagnostic accuracy and reducing the burden of manual data collection. The complete integration of AI, from data acquisition to automated anamnesis, marks a significant step toward scalable, intelligent, and patient-centered healthcare solutions.

2. Background

The background section is structured to first introduce the reader to fundamental concepts in clinical practice, including medical procedures, patient anamnesis, and the use of biomedical sensors. This foundation is intended to provide a basic understanding of how medical diagnoses are typically conducted. Building this context will help the reader comprehend how certain elements of the diagnostic process can be augmented or partially replaced by deep learning methods. Following this, section 2.2 will present the technical principles of deep learning that are essential for understanding the approaches developed and discussed throughout this thesis.

2.1. Medical Examinations

2.1.1. Fundus Images

Fundus images are high-resolution photographs that capture the detailed structure of the retina, the light-sensitive tissue located at the back of the eye [20]. These images are obtained using a specialized ophthalmic camera equipped with a flash illumination system. The process involves directing a controlled beam of light into the eye, which passes through the cornea, aqueous humor, lens, and vitreous body before reaching the retina [21]. The light illuminates the intraocular region, enhancing visibility of the retinal structures, including the optic disc, macula, blood vessels, and surrounding tissues [22, 23].

A schematic representation of the eye anatomy is provided in figure 2.1 to illustrate the passage of light and the imaging process. The light that interacts with the retinal surface is partially absorbed and partially reflected. The reflected light is then captured by the camera sensor, producing a detailed digital representation of the retina.

The quality and characteristics of fundus images can vary due to several factors. One key factor is the viewing angle of the camera, which may change depending on the specific examination technique, patient positioning, and ocular conditions [20, 21]. Variations in the angle and alignment can lead to differences in image perspective, illumination uniformity, and visibility of retinal features, potentially influencing the interpretation of retinal health [24]. Examining the surface area of the retina through fundus imaging provides valuable insights into a person's overall health, particularly the condition of the

ocular and systemic circulatory systems. The retina is a highly vascularized tissue, and changes in its structure can serve as indicators of underlying medical conditions [21, 22, 23].

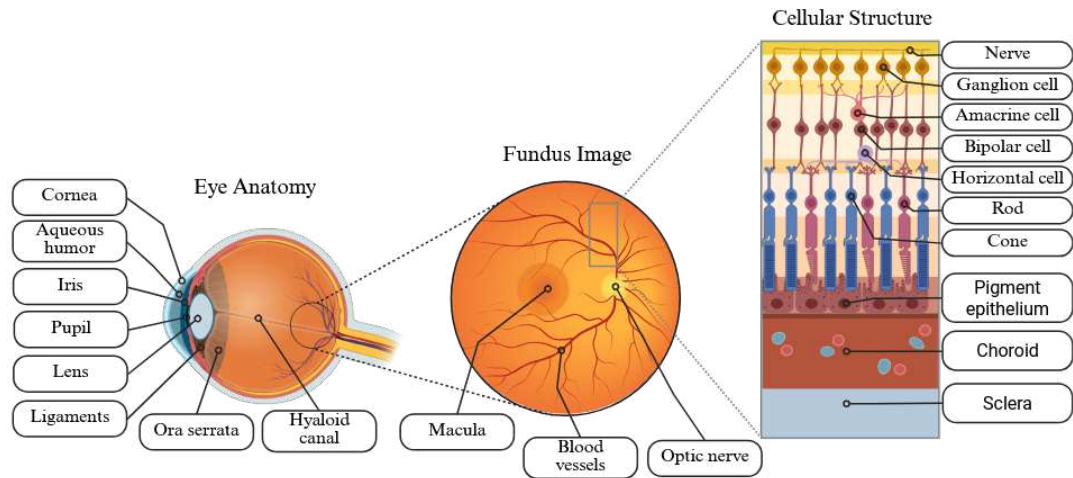


Figure 2.1.: The image provides a detailed illustration of the anatomical structure of the eye. In the process of fundus photography, light must first traverse the external ocular structures, including the cornea and lens, before entering the intraocular region. Once inside, the light reaches and illuminates the retinal structure, allowing for the visualization of key retinal components such as the optic disc, macula, and blood vessels. The image was created with BioRender.com.

2.1.2. Diagnostics using Fundus Images

Fundus images serve as crucial diagnostic tools by capturing detailed information about the structural and vascular integrity of the retina [21, 22, 23]. These images not only provide insights into the health status of the eye but also offer valuable indications of systemic diseases affecting the body [20, 25, 26, 27, 28, 29]. Structural abnormalities, vascular changes, and signs of cellular damage in different regions of the retina can signal the presence of various ocular and systemic conditions. Through detailed analysis of fundus images, healthcare professionals can detect, monitor, and assess the progression of these pathological changes, enabling early intervention and treatment.

Diabetic Retinopathy (DR), primarily associated with Type 1 and Type 2 diabetes, is one of the most prevalent retinal disorders detectable through fundus imaging [20, 25]. However, several other ocular diseases also contribute significantly to global vision impairment and blindness. Among these, cataracts, glaucoma, and Age-related Macular Degeneration (AMD) are particularly concerning, especially in aging populations. According to the WHO [30], the five leading causes of vision impairment worldwide are diabetic retinopathy, cataracts, glaucoma, AMD, and refractive errors. The prevalence of these conditions underscores the importance of early diagnosis, regular ophthalmic examinations, and timely medical or surgical interventions to prevent irreversible vision loss and promote ocular

health. Fundus photography plays a significant role in the early diagnosis of DR, glaucoma, and AMD.

AMD primarily affects the macula, the central portion of the retina responsible for detailed vision (see figure 2.1) [26, 27]. AMD is classified into neovascular AMD (wet) and geographical atrophic AMD (dry). The latter form is characterized by the gradual thinning of macular tissue and the accumulation of drusen, extracellular deposits beneath the retina visible through fundus photography. The neovascular form involves the abnormal growth of fragile blood vessels, which can lead to hemorrhages and rapid vision loss. Early detection of AMD is critical, as interventions such as anti-vascular endothelial growth factor therapy can slow disease progression and preserve vision [31]. The presence and distribution of drusen in fundus images serve as early indicators of AMD, aiding in diagnosis and disease monitoring.

Glaucoma refers to a group of neurodegenerative disorders characterized by progressive damage to the optic nerve, often associated with elevated intraocular pressure [28, 29]. This condition is particularly concerning because it develops gradually and remains asymptomatic in its early stages [28]. If left untreated, glaucoma leads to irreversible vision loss due to the gradual loss of retinal ganglion cells and damage to the optic nerve head and nerve fiber layers. The diagnosis of glaucoma often involves fundus imaging, with a focus on assessing the optic nerve head for structural abnormalities such as increased cupping, thinning of the neuroretinal rim, and retinal nerve fiber layer defects [29].

Figure 2.2 presents fundus images illustrating different pathological manifestations of these diseases. While fundus photography is an essential tool in ophthalmology, the accurate classification of these conditions remains a challenge, even for experienced specialists [32]. The complexity arises from subtle differences in retinal abnormalities, variations in disease progression, and overlapping features between multiple conditions. As a result, advanced image analysis techniques, including artificial intelligence-based diagnostic models, are increasingly being explored to enhance accuracy and reliability in detecting and classifying retinal diseases.

Given the non-invasive nature of fundus imaging and the capabilities offered by an AI-driven decision support system, the fundus camera was incorporated into the Virtual Doctor project. This integration is designed to enhance diagnostic accuracy while streamlining and accelerating the overall diagnostic workflow.

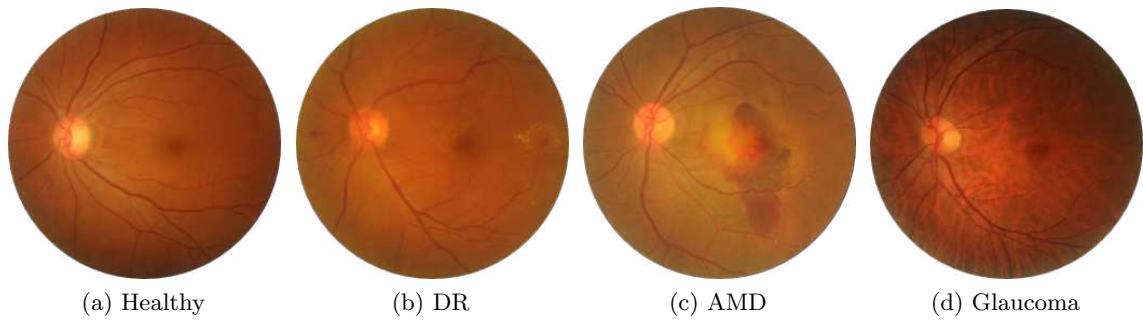


Figure 2.2.: The figure illustrates distinct pathological features that are characteristic of the three eye diseases presented. For reference, subfigure (a) displays a healthy fundus image to provide a visual comparison. The images were taken from the FIVES dataset [33].

2.1.3. Skin Lesion Examinations

Skin cancers represent the most commonly diagnosed category of cancers globally [34]. According to data from the World Health Organization, approximately 330,000 new cases of melanoma were diagnosed worldwide in the year 2022 alone. In the same year, nearly 60,000 individuals died as a result of skin cancer, highlighting its significant impact on global public health [35].

Early detection is crucial for improving patient outcomes, as timely diagnosis and intervention can greatly increase the effectiveness of treatment. Melanoma is the most lethal form of skin cancer, accounting for approximately 75% of all skin cancer-related deaths [36, 37, 38]. However, when melanoma is identified at an early, localized stage, the five-year survival rate exceeds 99% [39]. In contrast, if the disease progresses to advanced stages with metastasis, the five-year survival rate drops significantly to around 34% [39]. This underscores the importance of early screening and diagnosis in reducing mortality.

A standard skin cancer screening typically begins with a visual examination performed by a dermatologist. This examination of body tissue is conducted using either the naked eye, magnifying glass, or a more advanced tool known as a dermatoscope [40].

A dermatoscope is a specialized optical instrument designed to improve the visualization of skin lesions. It employs polarized light to enhance contrast and reduce surface reflections, allowing for a clearer view of subsurface skin structures. To achieve this, a corresponding polarization filter is placed before the camera, selectively reducing surface reflection while revealing underlying pigmentation patterns, vascular structures, and other morphological features critical for accurate diagnosis [41, 42]. Figure 2.3 illustrates the examination process and provides an example of a skin lesion recorded both with and without a polarization filter, demonstrating the improved visibility achieved through dermoscopic imaging.

If a dermatologist identifies a suspicious skin lesion during the initial visual examination, further diagnostic steps are required. In such cases, the physician may recommend an invasive procedure involving the excision of the affected skin tissue [43]. This removed tissue is then subjected to histopathological analysis through a biopsy, in which a pathologist examines the sample under a microscope to determine the presence of cancerous cells. This confirmatory diagnostic approach allows for precise classification of the lesion and guides subsequent treatment decisions.

A dermatoscope was integrated into the Virtual Doctor cabin to capture high-resolution, polarized images of skin lesions. Combined with AI analysis, this enables enhanced visualization, supports early detection, and improves diagnostic accuracy.

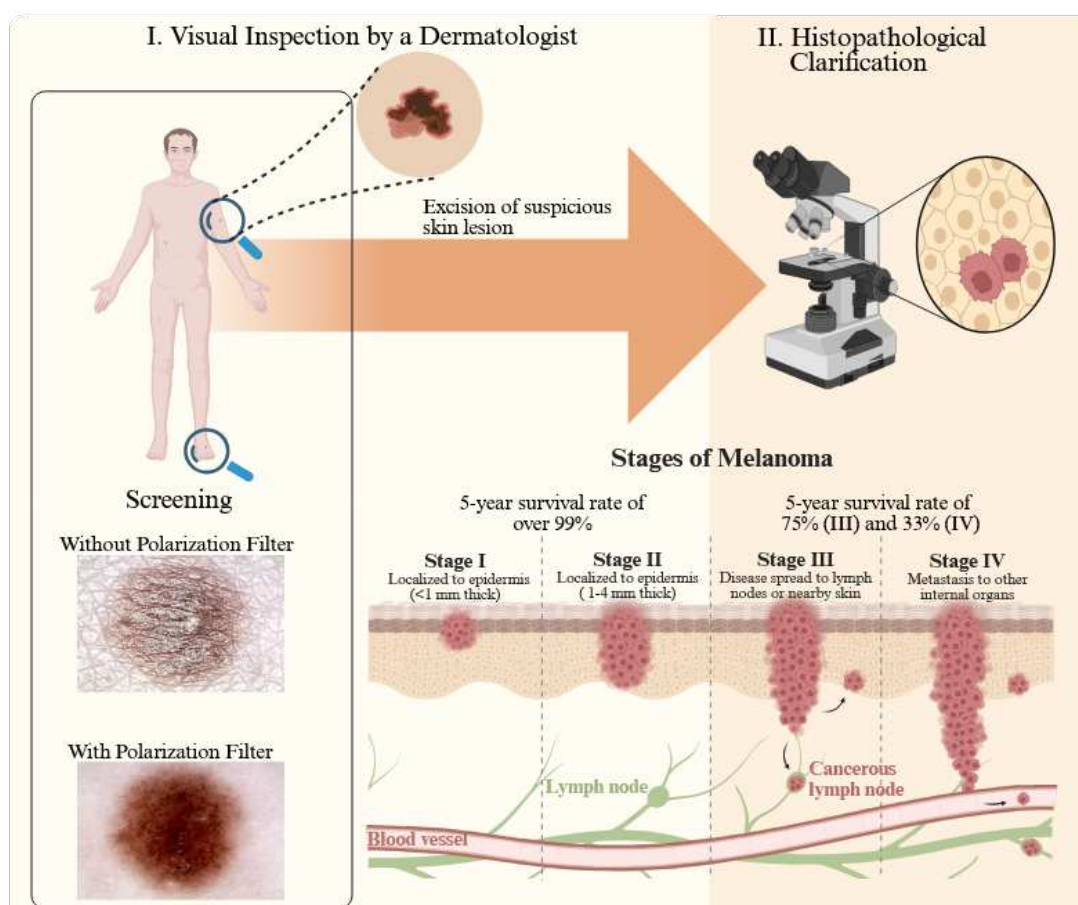


Figure 2.3.: The figure depicts the diagnostic workflow for skin cancer screening. The process begins with a visual examination of the skin, often supported by high-resolution optical tools such as a dermatoscope, which enhances tissue visibility through the use of polarization filters. If a lesion appears suspicious, a biopsy is performed by excising a sample of the affected tissue, which is then subjected to histopathological analysis in a laboratory. The prognosis, including the five-year survival rate, varies significantly depending on the stage at which melanoma is diagnosed. The illustration was created using BioRender.com and self-recorded skin lesions.

2.1.4. Type II Diabetes Mellitus

In contrast to type I diabetes, which is primarily associated with genetic predisposition and autoimmune mechanisms, Type 2 Diabetes Mellitus (T2DM) arises from a combination of impaired insulin secretion by pancreatic β -cells and the reduced responsiveness of insulin-sensitive tissues [44]. This condition is frequently driven by various cellular stressors, including chronic inflammation, endoplasmic reticulum stress, oxidative and metabolic stress, and amyloid deposition [45]. These factors can compromise β -cell function and may ultimately lead to a loss of islet structural integrity. Additionally, the presence of hyperglycemia and hyperlipidemia often exacerbates these stress responses, further promoting disease progression [46].

Multiple risk markers have been identified in relation to the development of type 2 diabetes mellitus. Notably, a geographical variation in T2DM prevalence has been observed across different populations [46]. Although the exact cause of this variation remains unclear, it is hypothesized that differences in lifestyle, socioeconomic conditions, and environmental exposures contribute significantly to this disparity [46]. In addition, genetic predisposition has been shown to play a substantial role in individual susceptibility to the disease [47]. Among all known risk factors, obesity has been consistently identified as the most influential [48]. As a result, Body Mass Index (BMI), which is calculated as weight divided by the square of height, is frequently used as a key indicator for assessing diabetes risk. Other lifestyle factors like healthy diet, alcohol consumption and smoking have also been shown to have an effect on the development of T2DM [48, 49]. table 2.1 provides an overview of selected risk markers associated with the development of T2DM and their proposed effects. Some of these effects exhibit non-linear relationships. For example, moderate alcohol consumption has been associated with a reduced risk of developing T2DM, while excessive consumption is linked to a significantly higher risk [49]. Other factors, such as ethnicity, remain inconclusive due to potential confounding variables like cultural lifestyle, which may independently influence both the likelihood of developing T2DM and the associated risk markers [46]. In addition to assessing established risk markers, physicians often inquire about specific clinical symptoms indicative of T2DM. These may include polyuria, polydipsia, and unexplained weight loss [50].

The Virtual Doctor cabin was equipped to retrieve key T2DM risk markers, including BMI, lifestyle factors, and clinical symptoms. By integrating these data with AI-driven analysis, the system enables a rapid assessment of individual diabetes risk, supporting early detection and informed clinical decision-making.

Table 2.1.: The table lists risk markers for T2DM derived from different publications with the proposed effect on the development of T2DM.

Risk Marker	Effect	References
Age	Higher risk with age	[48]
Biological sex	Different risk perception	[49, 51]
Obesity	Increased risk	[44, 48]
Waist Circumference	Increased risk	[48]
Diet	Healthy diet reduces risk	[48]
Alcohol Consumption	Non-linear relationship	[48, 49]
Smoking	Increased risk	[48]
Physical Inactivity	Increased risk	[48]
Family History	Genetic predisposition increases risk	[44, 45, 46, 47]
Ethnicity	Potential risk for certain groups (ambiguous)	[46]

2.1.5. Anamnesis and Measures of Well-Being

One of the most fundamental and widely used methods in medical practice is the conversation between the physician and the patient. This diagnostic interaction, known as anamnesis, plays a crucial role in extracting critical information regarding the patient’s medical history, symptoms, and potential risk factors. Through anamnesis, physicians aim to identify key risk markers that can help narrow down the range of possible conditions affecting the patient. Based on the findings from this initial assessment, the physician may either arrive at a benign diagnosis or determine the need for further diagnostic investigations [52].

Anamnesis is typically tailored to the patient’s specific complaints and symptoms. For instance, if a patient reports experiencing rapid vision loss, the physician will likely proceed with an ophthalmological evaluation, such as a fundus examination, to assess potential retinal abnormalities. This targeted approach ensures that diagnostic procedures are directed toward the most relevant medical concerns, improving efficiency and accuracy in disease detection.

Beyond symptom-oriented anamnesis, researchers have investigated broader strategies to assess general health status through standardized well-being metrics. These approaches aim to provide a more holistic view of an individual’s health by capturing both physical and mental well-being, along with factors such as quality of life and satisfaction with healthcare services. Tools like the General Health Rating Index have been developed to quantify these aspects, although they present inherent challenges [53, 54]. Comprehensive questionnaires risk becoming overly lengthy and burdensome for patients, whereas shorter forms may omit important health indicators. Moreover, patients’ perceptions of health and interactions with the healthcare system may affect their responses, introducing subjectivity into the results.

In clinical practice, it is common to use standardized health questionnaires to document

patient history [55]. However, these instruments often focus narrowly on specific symptoms or diseases and may overlook broader or unrelated health issues. While they serve as useful tools to streamline information collection, they may miss nuanced or emerging conditions.

To complement traditional anamnesis, non-invasive biomedical sensors have been increasingly implemented, especially for screening asymptomatic or mildly symptomatic conditions. During the COVID-19 pandemic, tools such as thermal imaging cameras and pulse oximeters were widely deployed in healthcare and public settings [56, 57]. These sensors can detect physiological markers like elevated skin temperature or reduced blood oxygen saturation, offering early signs of potential illness without requiring direct physical contact. This enables rapid, low-risk screening in high-throughput environments and reduces the burden on clinical staff. When integrated with anamnesis, these sensor-based approaches can enhance diagnostic precision while maintaining safety and efficiency.

The Virtual Doctor cabin uses a large language model to conduct automated anamnesis, collecting patient history, symptoms, and risk factors. Combined with sensor data, this enables rapid, structured assessments that support early detection and informed clinical decisions.

2.1.5.1. Pulse Oximetry

The pulse oximeter is a noninvasive medical device used to measure both pulse rate and the oxygen saturation level in the blood, commonly referred to as Peripheral Capillary Oxygen Saturation (SpO_2). It operates based on the principle of light absorption and relies on the fact that oxygenated hemoglobin and deoxygenated hemoglobin absorb light differently at specific wavelengths [58, 59].

To perform the measurement, the device uses two light-emitting diodes that produce light at distinct wavelengths. One diode emits red light with a wavelength of approximately 660nm, while the other emits infrared light at around 940nm [58]. As this light travels through or is reflected by body tissue, such as a fingertip or earlobe, the amount of absorbed light depends on the relative concentrations of oxygenated and deoxygenated hemoglobin in the blood.

A photodiode on the opposite side or in a reflective position detects the transmitted or reflected light. The device then analyzes the signals received at both wavelengths. Oxygenated hemoglobin absorbs more infrared light and less red light, whereas deoxygenated hemoglobin shows the opposite pattern [58]. By comparing the intensity of light absorption at the two wavelengths the oxygen saturation of the blood can be estimated.

The pulse rate is derived by examining the rhythmic fluctuations in light absorption, which are associated with the pulsing flow of blood through the capillaries during each heartbeat. Because the measurement process is based entirely on light absorption and detection, it enables continuous and real-time monitoring of vital physiological parameters without the

need for invasive procedures such as drawing blood samples. Figure 2.4 illustrates the principle of pulse oximetry for red light, showing how the light interacts with the tissue and the corresponding pulsation of the measured signal. For infrared light the time trace in figure 2.4 is shifted as the absorption of oxygenated hemoglobin is higher than that of deoxygenated hemoglobin.

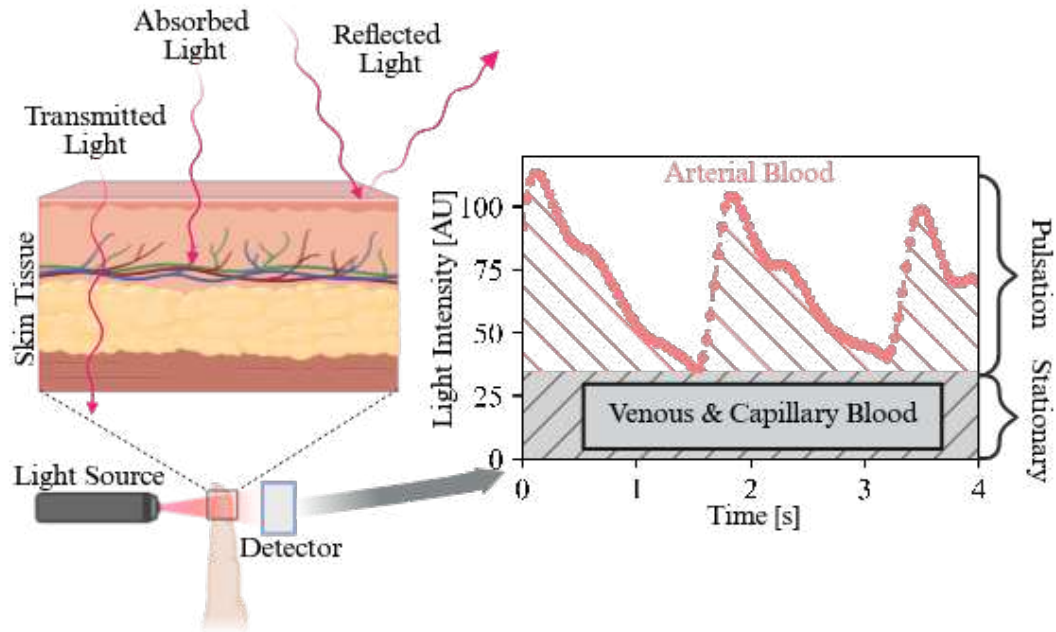


Figure 2.4.: The figure illustrates the principle of pulse oximetry, using the absorption characteristics of oxygenated and deoxygenated hemoglobin. Light of a specific wavelength (660nm for red light and 940nm for infrared light) is transmitted through the tissue and the amount of light absorbed is measured to determine the oxygen saturation level. The pulsation of the measured signal corresponds to the heartbeat and is caused by arterial blood flow. The illustration was created using BioRender.com, a self-recorded pulse oximetry signal using the Beurer PO80 pulse oximeter, and is adapted from [58].

From an anamnesis perspective, pulse oximeter data can offer important insights into a patient's cardiopulmonary function. For instance, persistently low oxygen saturation values ($<95\%$) may indicate underlying conditions such as chronic obstructive pulmonary disease, asthma, or early signs of respiratory infections like COVID-19 [60, 61]. Sudden drops in SpO_2 , especially when accompanied by symptoms such as shortness of breath or fatigue, may warrant further clinical investigation.

Additionally, the pulse rate measured by the device can reveal abnormalities such as bradycardia (unusually slow heart rate) or tachycardia (elevated heart rate), which may reflect cardiovascular dysfunction [62]. When integrated into an automated anamnesis system, pulse oximeter readings can help flag potential acute or chronic health issues and support early diagnostic decisions.

2.1.5.2. Thermal Camera

A thermal camera is a non-contact imaging device that detects infrared radiation emitted by objects and converts it into a visual representation known as a thermogram [63]. Unlike standard cameras that capture visible light, thermal cameras detect electromagnetic radiation in the infrared spectrum, typically within the wavelength range between $8\text{ }\mu\text{m}$ and $14\text{ }\mu\text{m}$ [64].

Applying Wien’s displacement law for the black body (a simplified assumption regarding the absorption and emission characteristics) [65] as an approximation for the thermal radiation of the human body with a temperature range between $36\text{ }^{\circ}\text{C}$ and $40\text{ }^{\circ}\text{C}$ corresponds to an estimated emitted peak wavelength range between $9.37\text{ }\mu\text{m}$ and $9.25\text{ }\mu\text{m}$. This range lies within the infrared spectrum and falls well within the detection capability of most thermal cameras [66].

The core of a thermal camera consists of a detection array made up of photodetectors [66]. These detectors correspond to the resolution of the camera and are responsible for capturing the incoming infrared light. The received signals are converted into temperature values using a pre-defined calibration curve [64]. As an example, the FLIR Lepton thermal camera [64] uses a lens system to direct the incoming infrared radiation onto a focal plane array. This array is composed of vanadium oxide microbolometer elements that sit atop an integrated readout circuit. The microbolometer elements experience a temperature change in response to incident infrared light flux, which leads to a measurable change in electrical resistance. This change is then translated into pixel values that form the final thermal image.

In medical applications, thermal cameras are used to visualize and quantify temperature distributions across the human body [66]. The skin temperature reflects underlying physiological processes, such as blood flow, inflammation, or metabolic activity [66]. By detecting even subtle asymmetries or localized temperature increases, thermal imaging can aid in identifying abnormal physiological states. The resulting thermographic data can be analyzed in real-time or processed further using AI algorithms to detect patterns associated with specific health conditions.

From an anamnesis perspective, thermal imaging can provide valuable diagnostic clues about a patient’s condition. For example, localized areas of elevated skin temperature may indicate inflammation, infection, or early signs of tissue damage [66]. Elevated temperatures on the face or forehead could suggest the presence of fever, a common symptom of infection or systemic inflammatory response [67].

2.1.6. Medical Studies

Medical studies are governed by strict regulatory frameworks and are subject to oversight by independent ethical review boards. Prior to initiating any study involving human participants, researchers are required to submit a comprehensive application to an ethics committee. This committee is responsible for evaluating the scientific rationale, ethical considerations, and legal compliance of the proposed research.

The submission typically includes a detailed study protocol, a participant consent form, and additional documentation depending on the nature and design of the study. The study protocol must clearly outline the scientific objectives, methodology, and expected outcomes. It must also include a thorough risk-benefit assessment, explaining how potential risks to participants are minimized and justified in relation to the anticipated benefits of the research.

In addition, the study documentation must provide a clear and precise description of the procedures involved, including the inclusion and exclusion criteria, the estimated number of participants, and the timeline of the study. Researchers must also submit a data management plan, detailing how personal and health-related data will be collected, stored, processed, and protected throughout the study. This includes outlining technical and organizational measures to ensure data privacy and compliance with national and international data protection laws.

In Germany, the legal framework for data protection in medical research is primarily governed by the General Data Protection Regulation, also known as *Datenschutz-Grundverordnung* (DSGVO). All research involving personal data must adhere to the provisions of this regulation. Participants must be thoroughly informed about their rights in relation to their data. These rights include the right to voluntarily participate or refuse participation without any disadvantage (Art. 7 DSGVO), the right to withdraw from the study at any time (Art. 7 DSGVO), the right to request the deletion of their personal data (Art. 17 DSGVO), and the right to file a complaint with a supervisory data protection authority (Art. 77 DSGVO).

Ethical approval is a fundamental prerequisite for conducting medical research involving human subjects. It ensures that participants are treated with dignity, that their rights and privacy are respected, and that the research adheres to both scientific integrity and legal standards.

2.2. Machine Learning

Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms capable of learning patterns from data. These algorithms are designed to recognize complex structures and relationships in datasets, enabling tasks such as prediction,

classification, and decision-making without relying on explicit rule-based programming [68, 69].

At its foundation, machine learning is formulated as an optimization problem. The learning process involves finding the optimal parameters of a model such that a predefined objective function, often referred to as the loss function, is minimized (or maximized). The formulation of this optimization problem, including the choice of the model and the definition of the loss function, is critical to capturing the underlying structure of the data.

In simple cases, such as linear regression with a least-squares loss function, the optimization problem has a closed-form solution. For instance, given a feature matrix $X \in \mathbb{R}^{n \times m}$, a corresponding ground truth vector $y \in \mathbb{R}^n$, and learnable weights $\omega \in \mathbb{R}^m$, the objective is to minimize the squared error between the model prediction and the true ground truth as described in equation (2.1). The optimal solution in this case is given by the Moore-Penrose pseudoinverse [68, 70].

$$\min_{\omega} \|X\omega - y\|_2^2 \quad (2.1)$$

However, for more complex models and data distributions, especially those involving nonlinearity, high dimensionality, or large-scale datasets, closed-form solutions are generally not available. In such scenarios, heuristic and iterative optimization techniques, such as gradient descent, are employed to approximate the optimal solution.

The effectiveness of a machine learning model depends not only on the expressiveness of the model architecture but also on the quality of the optimization process and the suitability of the loss function with respect to the learning objective. These components together define the framework within which the model learns from data.

The integration of machine learning into medicine has led to substantial advancements across diagnostic, prognostic, and therapeutic applications [71, 72, 73]. Machine learning models are capable of analyzing large volumes of medical data such as imaging, genetic information, and electronic health records with a level of speed and precision that surpasses manual analysis. This enables earlier disease detection, personalized treatment planning, and improved clinical decision-making. For instance, deep learning algorithms have shown expert-level performance in tasks like skin cancer classification, retinal disease detection, and automated anamnesis [74, 75, 76]. The ability to uncover hidden patterns in complex datasets makes machine learning an essential tool in modern healthcare, offering the potential to improve patient outcomes and optimize the efficiency of medical services.

2.2.1. Camera Calibration

Camera calibration is a fundamental step in translating image-plane measurements into real-world dimensions. This process is essential in medical applications where anthropometric

estimates, such as the waist-to-hip ratio, can be derived from camera recordings and used as risk indicators for various conditions, including Type 2 Diabetes Mellitus (see section 2.1.4).

The geometric relationship between a three-dimensional point in the world coordinate system and its corresponding projection onto the two-dimensional image plane is defined by a camera model that incorporates both extrinsic and intrinsic parameters. The extrinsic parameters represent the rotation matrix $R \in \mathbb{R}^{3 \times 3}$ and translation vector $T \in \mathbb{R}^3$, which describe the position and orientation of the camera relative to the world coordinate system. The intrinsic parameters, typically denoted as matrix $A \in \mathbb{R}^{3 \times 3}$, encode the internal characteristics of the camera, such as focal length, skew, and principal point offset [77, 78]. The projection of a point from the world coordinate system to the image plane is defined in equation (2.2). The projection matrix P is the composition of the R , T , and A .

$$c \cdot x_{image} = P(R, T, A) \cdot x_{world} \quad (2.2)$$

c describes a scaling factor, which can be determined by the third dimension component in real world coordinates. x_{image} is the projected image point and x_{world} the corresponding 3D point in world coordinates. This transformation can be reformulated into a linear system of equations and rewritten as an optimization problem in the form of equation (2.3).

$$\min_{p, \|p\|_2^2=1} \|Kp\|_2^2 \quad (2.3)$$

Here, p is the vectorized form of the projection matrix P , and K is a design matrix constructed from known 3D-2D point correspondences. The solution to this system can be analytically found using Singular Value Decomposition (SVD) [79]. The optimal p corresponds to the right singular vector associated with the smallest singular value of K , providing a closed-form approximation of the projection matrix. A detailed derivation of equation (2.3) is provided in appendix B.1.

Before applying this transformation, however, it is necessary to correct for optical distortions introduced by the camera lens. These distortions, if unaccounted for, can lead to significant inaccuracies in real-world coordinate estimation.

2.2.1.1. Radial Distortion

Radial distortion arises due to the geometry of the lens system, particularly from the shape and position of the aperture ring [77, 78, 80]. It causes straight lines in the real world to appear curved in the image. This type of distortion can be corrected by assuming a polynomial correction term as given in equation (2.4).

$$r_u = r_d \cdot (1 + \delta r) = r_d \left(1 + \sum_{i=0}^{\infty} \kappa_i r_d^i \right) \quad (2.4)$$

r_d is the distorted radius, r_u is the corrected (undistorted) radius, and the κ_i variables are distortion coefficients, which have to be determined. Typically, only the second- and fourth-order terms of the distortion coefficients are retained in practical implementations, as higher-order terms have shown limited improvement in empirical studies [80].

2.2.1.2. Tangential Distortion

Tangential distortion results from misalignment between the lens and the image sensor. It leads to a shift of image points away from their ideal positions [81]. The correction model can be geometrically derived and is defined by equation (2.5).

$$\begin{aligned} x_u &= x_d + [2p_1 x_d y_d + p_2 (r^2 + 2x_d^2)] \\ y_u &= y_d + [p_1 (r^2 + 2y_d^2) + 2p_2 x_d y_d] \end{aligned} \quad (2.5)$$

p_1 and p_2 are tangential distortion coefficients, and r is the radial distance from the image center. Due to the nonlinear nature of these distortion correction models, a closed-form solution is not available. Instead, iterative optimization algorithms such as Levenberg-Marquardt are employed to estimate the camera parameters [82]. This method is implemented in widely used libraries such as OpenCV, which provide robust calibration pipelines for both intrinsic and extrinsic parameter estimation [83].

Accurate camera calibration is a critical step for extracting meaningful quantitative measures from image data, particularly in clinical environments where precise measurements can influence diagnostic and treatment decisions.

2.2.2. Deep Learning

In the previous section 2.2.1, it was shown that from physical properties of systems an analytical (see equation (2.2)) or heuristic model (see equation (2.4)) can be derived and solved either analytically or with heuristic optimization approaches. In the same fashion often model definition is a complex task and often an analytic model, while more powerful due to the interpretable nature, is not required for the problem solution.

In this context deep learning, a subfield of artificial intelligence and machine learning, focuses on approximating solutions to data driven tasks by constructing compositions of functions with learnable parameters [69, 84]. These functions are not explicitly problem specific, but may be motivated by the problem field [85]. These functions are motivated by

the human brain mechanism and hence described as neural networks which are composed of multiple layers of interconnected nodes, or neurons. Each “neuron” is defined by a learned nonlinear function to transform the input data. The term “deep” refers to the presence of many such composition layers, which allow the network to gradually extract more abstract and meaningful features from raw input data. The analogy between the network architecture in deep learning and the human brain mechanism is depicted in the figures 2.5a and 2.5b.

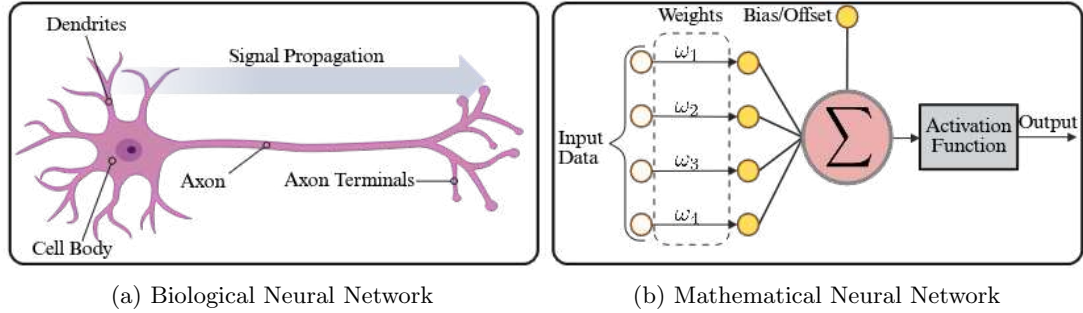


Figure 2.5.: The figure compares the communication pathways of biological neurons with those of artificial neural networks in machine learning. Subfigure (a) illustrates the biological process, where incoming signals are received by the dendrites and integrated within the neuron’s cell body. If the combined signal exceeds a certain threshold, an action potential is triggered and transmitted along the axon to communicate with other neurons via synapses. Subfigure (b) presents the analogous process in an artificial neural network. Here, each input feature is multiplied by an associated weight, which represents a learnable parameter. The weighted inputs are summed together with an additional bias term, which serves as an adjustable offset. This aggregated value is then passed through an activation function, which introduces non-linearity into the model. The resulting output can subsequently serve as the input to the next layer, allowing the network to build increasingly complex feature representations. The figure was created using BioRender.com and is adapted from [86].

There are several key types of neural network architectures commonly used in deep learning. Especially, the Multilayer Perceptron (MLP) for feature data, Convolutional Neural Network (CNN) for image tasks and transformer architecture for Natural Language Processing (NLP) [84, 87, 88]. More recently, transformer models have become highly influential due to their ability to capture long-range dependencies and their effectiveness across a wide range of tasks [89]. In the medical field, deep learning has shown promising results in areas such as image-based diagnostics, disease classification, segmentation of medical scans, and prediction of patient outcomes [90]. These models can be trained on large sets of annotated medical data and have achieved accuracy levels that, in some cases, are comparable to or better than human experts [74, 75]. Despite these successes, challenges remain in terms of data quality, model interpretability, and robustness [91]. These challenges are particularly important in clinical environments where accuracy, trust,

and safety are critical.

2.2.2.1. Gradient Descent and Optimization Techniques

Gradient descent is one of the fundamental optimization techniques used in deep learning. It is based on a heuristic approach that involves iteratively updating the learnable parameters of a model in order to minimize a given loss function [68, 69]. Each update step is determined by computing the gradient of the loss function with respect to the parameters. This gradient points in the direction of the steepest ascent, and its negative is used to adjust the parameters towards a local minimum. The parameter update rule for each iteration, often referred to as an epoch, is described in equation (2.6), where ω denotes the learnable parameters, $\mathcal{L}$ is the loss function, M is the model, y is the ground truth, and τ represents the learning rate.

$$\omega_{t+1} = \omega_t - \tau \cdot \nabla_{\omega} \mathcal{L}(M(\omega_t), y) \quad (2.6)$$

The learning rate plays a crucial role in the behavior of the optimization process. A large learning rate can cause the updates to overshoot the optimal solution, leading to instability or divergence. In contrast, a very small learning rate can make the optimization process extremely slow and inefficient. To address this, several techniques have been developed to adjust the learning rate during training. One common method is learning rate reduction on plateau, where the learning rate is decreased if the loss remains unchanged over a specified number of iterations. Another strategy is learning rate annealing, where the learning rate is reduced progressively based on a predetermined schedule depending on the training progress [92].

A common variant of gradient descent used in practice is Stochastic Gradient Descent (SGD). In contrast to standard or batch gradient descent, which uses the entire training dataset to compute the gradient at each step, SGD estimates the gradient using only a single sample or a small subset of data (a mini-batch). This leads to faster and more frequent updates, which can help escape local minima and speed up convergence. However, it introduces more variance in the optimization trajectory, which may lead to fluctuations around the minimum. In many deep learning applications, SGD and its mini-batch version have become the default approach due to their efficiency and scalability on large datasets [69].

In addition to these basic forms, more advanced optimization methods have been introduced to improve convergence speed and stability. Second-order approaches, such as the Newton method and the Levenberg-Marquardt algorithm, utilize information about the curvature of the loss surface by incorporating second-order derivatives or approximations [82]. These

methods are particularly useful in cases where simple gradient-based methods converge too slowly, but they are computationally more expensive and less practical for large-scale deep learning problems.

Another highly influential method is Adaptive Moment Estimation (Adam). This technique computes individual adaptive learning rates for each parameter based on estimates of both the first moment (the mean of the gradients) and the second moment (the uncentered variance). By adjusting the update size for each parameter individually, Adam improves learning stability and accelerates convergence, especially in sparse or noisy environments. It has become one of the most widely used optimizers in modern deep learning frameworks [92].

It is important to recognize that in most real-world scenarios, especially those involving non-convex loss functions, gradient descent and its variants do not guarantee convergence to a global minimum. Instead, these methods typically find a local minimum that is good enough for practical purposes. Despite their heuristic nature, gradient-based optimization techniques, including SGD and Adam, have shown exceptional performance across a wide range of machine learning tasks and remain at the core of deep learning model training [93].

2.2.2.2. Multilayer Perceptron and Activation Functions

Multilayer perceptrons are a foundational class of feedforward neural networks used in many deep learning applications. In an MLP, each element of the input vector is connected to subsequent layers through learnable parameters known as weights. These weights scale the inputs, which are then summed and passed through a non-linear transformation called an activation function. A single such transformation represents one layer of the network. By stacking multiple layers, an MLP is able to learn increasingly complex and abstract representations of the input data.

Each hidden layer in the MLP contributes to the model’s ability to approximate non-linear functions. Without non-linear activation functions between layers, the model would collapse into a simple linear transformation, regardless of its depth. Therefore, the choice and design of activation functions are critical to the performance and expressiveness of deep neural networks.

The selection of activation functions remains an open area of research, with the optimal choice often depending on the specific characteristics of the task at hand [85]. Among the most commonly used is the Rectified Linear Unit (ReLU), which is defined by the function (2.7).

$$f(x) = \max(0, x) \tag{2.7}$$

ReLUs introduce sparsity in the activations and are computationally efficient. However,

they can suffer from issues such as gradient discontinuity and the "dying ReLU" problem, where neurons project to the zero space and hence stop learning during training [94].

To address these limitations, alternative activation functions have been proposed. The Gaussian Error Linear Unit (GELU) and the Softplus function are two examples that provide smooth, differentiable transitions across the input space [95, 96]. These functions tend to mitigate gradient vanishing problems by ensuring that gradients remain informative throughout training. Nevertheless, they are more computationally intensive due to their increased functional complexity, which may impact training speed, especially in large models or real-time applications.

At the output layer, the activation function is typically chosen based on the nature of the prediction task. For binary classification, the sigmoid function is often used. It maps real-valued inputs to the interval between 0 and 1, effectively modeling a probability. For multi-class classification problems, the softmax function is often applied to produce a normalized probability distribution over multiple output classes. Both of these output activation functions ensure that the final predictions conform to the properties of probability distributions, where outputs are bounded between 0 and 1 and the sum of all predicted probabilities equals one [68].

MLPs can serve as function approximators to estimate the probability of disease occurrence based on diverse physiological and demographic features, such as age, height, and weight. In the context of T2DM, for instance, predictive models derived from MLPs can be utilized to identify individuals at elevated risk. Such predictions provide physicians with an additional layer of decision support by highlighting potential risk indications and thereby facilitating earlier preventive interventions.

2.2.2.3. Convolutional Neural Networks

Convolutional Neural Networks are a specialized class of deep learning models particularly effective in analyzing structured grid-like data such as images. The core operation in CNNs is the convolution, which allows the model to extract local features by applying a set of learnable filters to the input data. For image data, this operation is defined by summing the pixel values within a localized image patch, weighted by a set of learnable parameters (the convolutional kernel K). The operation is mathematically described in equation (2.8) in which the convolution with the kernel K is performed over an image patch at pixel point I_{ijk} defined by the summation over m and n at each color channel c . The resulting output depth i is defined by the number of filters applied to the input image.

$$I'_{ijk} = \sum_c \sum_m \sum_n K(c, m, n) I_{c(j+m)(k+n)} \quad (2.8)$$

This operation is conceptually equivalent to the application of analytical filters in classical image processing, such as edge or corner detectors. In figure 2.6, a corner detection filter $B \in \mathbb{R}^{4 \times 4}$ (block matrix) defined in equation (2.9) is applied to a chessboard image to identify structural keypoints. The extrema in the filtered output, represented as local minima (blue) and maxima (red), correspond to the physical corners of the chessboard pattern. Within CNNs, analogous filters are not predefined but are instead learned directly from data through optimization procedures such as gradient descent. These learned kernels enable the network to hierarchically extract salient features and regions of interest that support downstream tasks, including classification, detection, and segmentation.

$$B = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad (2.9)$$

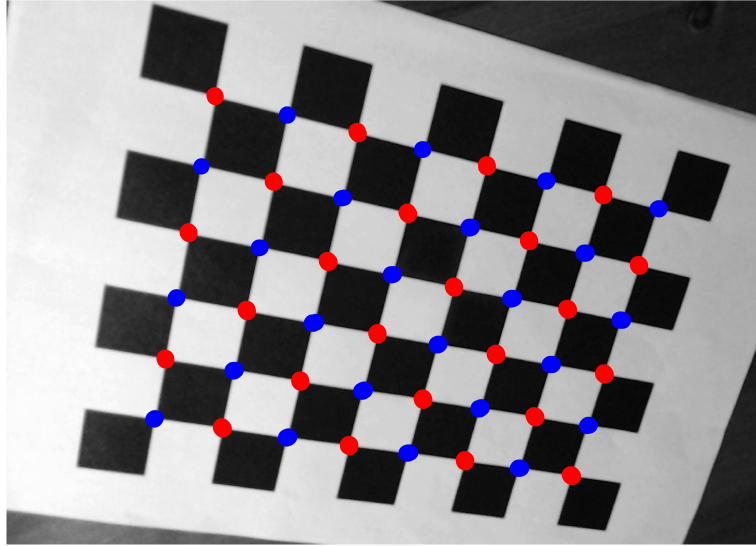


Figure 2.6.: The figure demonstrates the application of convolutional filters to images in a CNN. A chessboard pattern was captured using a Logitech C920 webcam. The filter defined in equation (2.9) was applied to the image, and local maxima (red) and minima (blue) were identified. Detection thresholds were set empirically at 1 for maxima and -1.1 for minima.

One of the key challenges in image analysis is the high dimensionality of the input data, particularly for high-resolution images. This increases computational demands and introduces the risk of overfitting due to the large number of parameters. To mitigate these issues, CNNs employ pooling layers that reduce the spatial dimensions of feature maps. Pooling operations, such as max pooling or average pooling, summarize local regions of the feature map by capturing the most relevant information, thus reducing data redundancy and the number of parameters to be learned [97].

CNNs owing to their inherent ability to detect and represent spatial hierarchies of features, have emerged as powerful tools for medical image analysis. Their capacity to recognize distinct morphological patterns makes them particularly valuable in diagnostic applications such as cancer detection in radiological imaging. Furthermore, due to their adaptability and diagnostic robustness, CNNs are well-suited to assist clinicians in identifying pathologically suspicious regions, for instance, in retinal fundus photographs. Beyond localization tasks, CNNs can also provide probabilistic diagnostic tendencies, such as in the assessment of skin lesions, thereby alerting physicians to potentially malignant structures and supporting informed clinical decision-making

2.2.2.4. Normalization Approaches

Normalization is a fundamental preprocessing technique in machine learning, aimed at scaling features to a comparable value range. Common approaches include min-max normalization, which linearly maps feature values to a specified interval (often $[0,1]$), and standardization, which centers each feature by subtracting its mean and scales it by dividing by its standard deviation. The latter approach has the advantage of utilizing statistical moments, thereby producing features with zero mean and unit variance, ensuring both comparable ranges and similar spread magnitudes.

In the context of neural networks, training is often performed using batch gradient descent, a variant of gradient descent in which weight updates are computed over small subsets of the training data. While this approach reduces memory requirements and can improve optimization stability, it may introduce fluctuations in the distribution of activations across network layers during training. This phenomenon, known as internal covariate shift, arises when the variance and mean of intermediate layer outputs change significantly between training steps [98].

To mitigate internal covariate shift, batch normalization is applied within the layers of the network [98]. This technique normalizes intermediate activations over each mini-batch, typically followed by a learnable scaling and shifting step to preserve representational flexibility. Empirically, batch normalization has been shown to accelerate convergence and improve training stability. The mathematical formulation of batch normalization is presented in equation (2.10), the transformed input features $\hat{x}^k$, with $k \in \{1, \dots, N\}$ and N being the number of features, are defined by the subtraction of the mean $\mathbb{E}[x^k]$ and division by the standard deviation $\sqrt{\mathbb{V}[x^k]}$ of the batch feature data x^k . The learnable parameters γ^k and β^k are used to scale and shift the normalized features, respectively.

$$\hat{x}^k = \frac{x^k - \mathbb{E}[x^k]}{\sqrt{\mathbb{V}[x^k]}} \cdot \gamma^k + \beta^k \quad (2.10)$$

Beyond batch normalization, alternative normalization strategies have been developed to address specific architectural or task-related constraints. Examples include layer normalization [99], instance normalization [100], and group normalization [101]. These methods share a similar mathematical structure to equation (2.10) but differ in the dimensions over which normalization statistics are computed, enabling adaptation to various data modalities and network designs.

2.2.2.5. Transformer

Transformer architectures were introduced in 2017 with the seminal publication “Attention Is All You Need” [88]. They are founded on the principle of learning contextual relationships between data elements through algebraic operations, primarily leveraging attention mechanisms to estimate similarity between representations. Although initially developed for NLP tasks, their versatility has since enabled their adoption across a wide range of domains such as computer vision [102].

In NLP applications, input text is first tokenized, i.e. transformed into a sequence of numerical representations. A basic tokenization approach assigns a unique integer to each word. However, this method leads to an overpopulation of possible representations for the same meaning (e.g. “do not”, “does not”, “don’t”, “doesn’t”). More sophisticated approaches, such as Byte Pair Encoding used in GPT-2 [103], convert words into UTF-8 bytes, thereby improving representation flexibility and enabling the handling of rare or unseen words.

The resulting token indices are then mapped to dense embedding vectors, which are learnable parameters. To preserve information about word order, a property absent from purely embedding-based representations, a positional encoding or positional embedding is added. The original transformer model employed sinusoidal positional encodings [88], whereas architectures such as GPT-2 [103] use learned positional embeddings.

The core operation of the transformer is multi-head self-attention. In this process, the input embeddings are linearly projected into three distinct sets of Query, Key, Value vectors. These are then partitioned into multiple attention heads, with the embedding dimension evenly divided among them. Within each head, attention scores are computed as the scaled dot product between queries and keys, followed by the application of a softmax function to obtain normalized weights. These weights are used to combine the corresponding values, producing the head outputs, which are then concatenated and linearly transformed.

In the encoder block of the transformer, the queries, keys, and values are derived from the same input sequence. In contrast, within the decoder block, self-attention is applied both to the decoder’s own inputs and across encoder–decoder attention, where the queries originate from the decoder and the keys and values come from the encoder. Each attention module

is followed by a feed-forward neural network (often implemented as an MLP) and residual connections with layer normalization, forming the building blocks of the architecture.

Modern transformer-based models, such as GPT and its successors, utilize only the decoder stack, optimized for autoregressive text generation. This architectural simplification retains the multi-head self-attention and feed-forward sublayers while omitting the encoder component.

Transformer models achieve state-of-the-art performance in natural language processing tasks due to their self-attention mechanism, which enables them to model long-range dependencies and contextual relationships across entire input sequences. This capability allows transformers to dynamically weigh the relevance of each token in a sequence with respect to all other tokens, facilitating richer and more flexible representations than traditional recurrent or feed-forward architectures. Empirical studies have demonstrated their superiority in tasks such as machine translation and question answering [88, 104].

Given that humans may occasionally overlook or forget critical information, transformer-based diagnostic approaches have emerged as a particularly promising tool. For instance, the anamnesis process can be augmented and supported through the deployment of a transformer-trained anamnesis system.

2.2.2.6. Data Augmentation

Data augmentation is a widely employed strategy in machine learning to enhance model robustness and improve generalization performance [105]. It involves applying systematic transformations to training data to encourage the model to learn invariances that are desirable for the target task. In image-based applications, common transformations include geometric modifications such as rotations, translations, scaling, and affine transformations, as well as photometric adjustments such as changes in brightness, contrast, or color balance. The introduction of controlled noise, such as Gaussian noise, can further improve model stability when encountering previously unseen or noisy inputs [106].

Figure 2.7 presents a collection of geometric (subfigures b-c) and non-geometric (subfigures e-h) data augmentation techniques applied to a skin lesion image. Histogram equalization as shown in subfigure (h) can also be used as a contrast enhancing preprocessing step. The application of augmentation techniques requires careful consideration, as it may introduce undesired noise into the data distribution, potentially impeding effective model optimization.

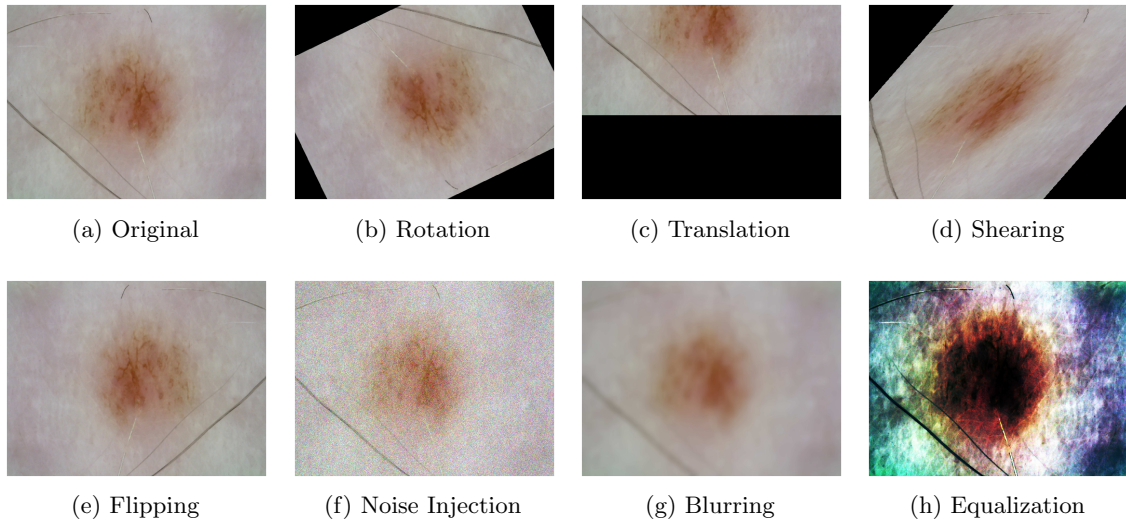


Figure 2.7.: The figure illustrates different augmentation techniques for images. The subfigures (b), (c), and (d) are geometric and subfigures (e) to (h) are non geometrical. The noise injection and blurring are gaussian based (gaussian kernel). The skin lesion was recorded using the Dinolite Dermatoscope.

Beyond improving invariance, augmentation can also serve as a tool for addressing domain shift. This issue arises when training and testing data differ systematically, which is common in medical imaging where variations in diagnostic devices, acquisition protocols, viewing angles, and image resolution can lead to dataset-specific biases. Without mitigation, such biases may cause the model to overfit to a particular imaging modality or protocol, thereby reducing its ability to generalize across clinical settings. Augmentation strategies that simulate these variations can help reduce such biases and improve cross-domain performance [107].

Data augmentation is frequently combined with upsampling in scenarios with limited datasets or class imbalance. In such cases, minority classes are often replicated to match the prevalence of majority classes. However, naive duplication risks overfitting to the repeated samples. Augmenting the upsampled data, by introducing transformations or controlled noise, can help diversify the replicated samples and reduce the likelihood of overfitting [108]. A practical example is Gaussian noise upsampling, where Gaussian-distributed noise is added to the upsampled images [109]. This alters the data distribution in a controlled manner, increasing sample diversity and potentially improving the model's ability to generalize.

2.2.2.7. Evaluation Metrics

In medical applications, the misinterpretation of machine learning results can have severe, potentially fatal consequences [110]. Consequently, careful selection, calculation, and

reporting of evaluation metrics is essential. Metrics can be chosen in ways that present an overly favorable picture of model performance, underscoring the need for methodological rigor in evaluation [71].

A robust evaluation protocol generally involves holding out a portion of the dataset exclusively for final testing, with the exact proportion depending on the size of the available data [99]. This test set must remain completely unseen during model training and hyperparameter tuning. Ideally, an entirely independent test set from a separate source is used, although such data is often unavailable in practice.

One commonly reported metric in classification tasks is accuracy, due to its intuitive interpretability. Accuracy is defined as the ratio of correctly classified samples to the total number of samples. However, in the presence of class imbalance, accuracy can be misleading, as a model biased toward the majority class may achieve high accuracy without true discriminative capability. To obtain a more nuanced performance assessment, additional metrics derived from the confusion matrix should be considered. The confusion matrix consists of the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) counts, which can be combined to compute more informative measures. For example the F1 score is defined as the harmonic mean of precision and recall as described in equation (2.11) [111].

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1 Score} &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{2.11}$$

The Matthews Correlation Coefficient (MCC), as defined in equation (2.12), constitutes a balanced performance metric that remains informative even in the presence of class imbalance.

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{2.12}$$

Another widely adopted performance indicator is the Receiver Operating Characteristic Area Under the Curve (ROC-AUC). It is obtained by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) for different classification thresholds. The AUC is the integral under this curve and is generally more robust to class imbalance than accuracy.

Even with a hold-out test set, sampling variability, such as the inclusion of outliers, can artificially inflate or deflate performance estimates. To address this, k-fold cross-validation

is employed on the training data to assess statistical consistency. In k-fold cross-validation, the dataset is partitioned into k equally sized subsets (folds). The model is trained k times, each time using (k-1) folds for training and the remaining fold for validation, ensuring that every fold serves as the validation set exactly once [111].

When comparing multiple models, statistical significance tests, such as the Wilcoxon signed-rank test, can be applied to cross-validation results to determine whether performance differences are statistically meaningful [111]. This level of rigor is crucial in medical applications, where overestimated model performance can lead to inappropriate clinical decisions.

2.2.2.8. Model Calibration

Once a predictive model has been trained, its raw output scores are often not directly interpretable in a probabilistic sense. This is particularly problematic in the medical domain, where physicians and clinical practitioners require transparent and calibrated probabilities rather than abstract model scores in order to make informed decisions. Without proper calibration, the outputs of machine learning models may either overestimate or underestimate the true likelihood of an outcome, potentially leading to suboptimal or even harmful clinical decisions. Consequently, calibration methods are employed to transform raw model scores into probabilistic values that more accurately reflect the underlying class distributions [112].

A variety of calibration techniques exist, each with distinct assumptions and methodological underpinnings. One widely applied family of approaches is binning-based calibration, such as Bayesian binning, in which the score space is partitioned into intervals, and empirical class probabilities are estimated within each bin. This method allows the raw model outputs to be mapped into probability estimates that are more closely aligned with observed frequencies [112].

Another advanced method is GUESS, which provides a more flexible probabilistic calibration strategy. Instead of partitioning the score space into bins, GUESS operates by fitting a Probability Density Function (PDF) to the score distribution of each class. Once the class-specific PDFs are estimated, Bayes' theorem is applied to calculate the posterior probability of a class given a model score. In this way, GUESS recalibrates the raw outputs into well-grounded probabilistic predictions that can be directly interpreted. This approach is particularly useful in settings where class imbalance or limited training data would otherwise distort the calibration process.

Calibration is therefore an essential post-processing step in predictive modeling, bridging the gap between abstract machine learning outputs and clinically interpretable probabilities. By applying methods such as Bayesian binning or GUESS, the predictive utility of models

can be significantly improved, ultimately supporting more reliable and trustworthy medical decision support.

3. Materials and Methods

The Virtual Doctor project is organized into several closely interconnected components, each addressing a distinct aspect of the system’s design, development, or application. Together, these components form an integrated framework for AI-assisted healthcare, combining hardware, software, decision support, and patient interaction.

The first part of the project focuses on the Virtual Doctor Diagnostic Cabin. This includes the design and development of the hardware, such as a medical cabin equipped with biomedical sensors capable of autonomously collecting patient data. The cabin’s components are interconnected through a distributed systems architecture, allowing secure remote access and monitoring by healthcare professionals. A structured database stores collected data, while software interfaces manage sensor communication, patient interaction, and system control, providing a functional and secure environment for autonomous clinical operations.

Building on this infrastructure, the project incorporates deep learning–based decision support. Advanced neural network architectures were developed for the analysis of medical datasets, including a novel adaptive normalization layer that dynamically adjusts to input data distributions for more robust learning. These models were applied to clinical use cases such as the analysis of retinal fundus images. Here, attention-based deep learning models not only accurately classify pathological findings but also localize them within the images, enhancing interpretability and clinical relevance.

Complementing the hardware and decision support, the Virtual Doctor Avatar facilitates patient interaction through a medical chatbot capable of performing automated anamnesis. The chatbot uses adaptive questioning to capture relevant medical history, storing information in a structured database. It is built on a LLM fine-tuned on a synthetic medical dataset to ensure medical domain accuracy. A graphical avatar provides a virtual physician interface, making interactions more engaging and human-like.

Finally, all project components were developed under the oversight of ethical approvals, ensuring patient safety, privacy, and compliance with medical research standards.

Together, these elements form a cohesive ecosystem, integrating hardware, deep learning decision support, and natural language–based patient interaction into the Virtual Doctor framework, demonstrating a comprehensive approach to modern AI-assisted healthcare.

3.1. Concept of the Virtual Doctor Diagnostic Cabin

The overarching objective of the Virtual Doctor cabin was the conception and implementation of a diagnostic environment capable of facilitating semi-supervised medical data acquisition while simultaneously providing decision support through advanced computational methods. Central to this approach was the integration of deep learning models into the cabin's software architecture, thereby enabling the automated processing of multimodal patient data. This data encompassed a wide range of inputs, including physiological sensor measurements, structured questionnaire responses, and medical imaging modalities. The embedded models were designed to analyze these inputs in real time and generate clinically meaningful outputs in the form of probabilistic risk assessments and diagnostic recommendations. Importantly, the system did not aim to replace physician expertise but rather to augment clinical decision-making by presenting model-derived insights through an interactive user interface, thereby ensuring that physicians could incorporate algorithmic guidance into their evaluations in a transparent and supportive manner.

The physical realization of the Virtual Doctor cabin was grounded in careful design considerations. The cabin's structure was developed using the Computer Aided Design (CAD) software FreeCAD 0.21.2, which allowed for precise planning of spatial dimensions, ergonomics, and hardware placement. A visual representation of the finalized design is provided in the appendix C.1. Following the digital prototyping phase, the required hardware components, such as biometric sensors, cameras, and computational units, were procured and systematically assembled.

3.1.1. Hardware Design

The three-dimensional model of the Virtual Doctor cabin served as the foundation for refining the cabin's spatial dimensions and ensuring ergonomic feasibility for patients and physicians alike. Based on iterative adjustments, the final cabin dimensions were set to 1.5 m in width, 1.8 m in length, and 2.4 m in height. The increased height was a functional requirement to accommodate an integrated sliding door mechanism, which necessitated a change in the construction material from wood to reinforced metal beams in order to provide adequate structural stability. The cabin floor was designed with a recessed square section to allow for the seamless integration of a custom-built weighing scale, ensuring unobtrusive operation during patient assessments.

One of the cabin walls was reinforced with an additional wooden support beam to provide the necessary stability for mounting the primary display screen, which serves as the central interaction interface. On this same wall, both a thermal imaging camera and a conventional RGB camera were installed, enabling multimodal data acquisition for diagnostic purposes. Within the interior of the cabin, a work table was positioned to host essential diagnostic peripherals, including a microphone, a keyboard, a pulse oximeter, and a dermatoscope.

These devices collectively support a variety of diagnostic processes and patient interactions. A comprehensive overview of all construction materials and their associated costs is provided in appendix C.3.

Figure 3.1 presents representative recordings from the various hardware components integrated into the Virtual Doctor cabin. All acquired data streams are processed in real time and subsequently stored within a local database, ensuring both secure storage and immediate accessibility for downstream analysis and decision-support tasks. Particular emphasis was placed on data privacy and patient confidentiality during the design of the data acquisition and storage pipeline. To minimize the risk of identifying patients through visual information, the system does not retain raw images from the RGB camera. Instead, only the processed silhouette or outline representations of the recorded individual are preserved.

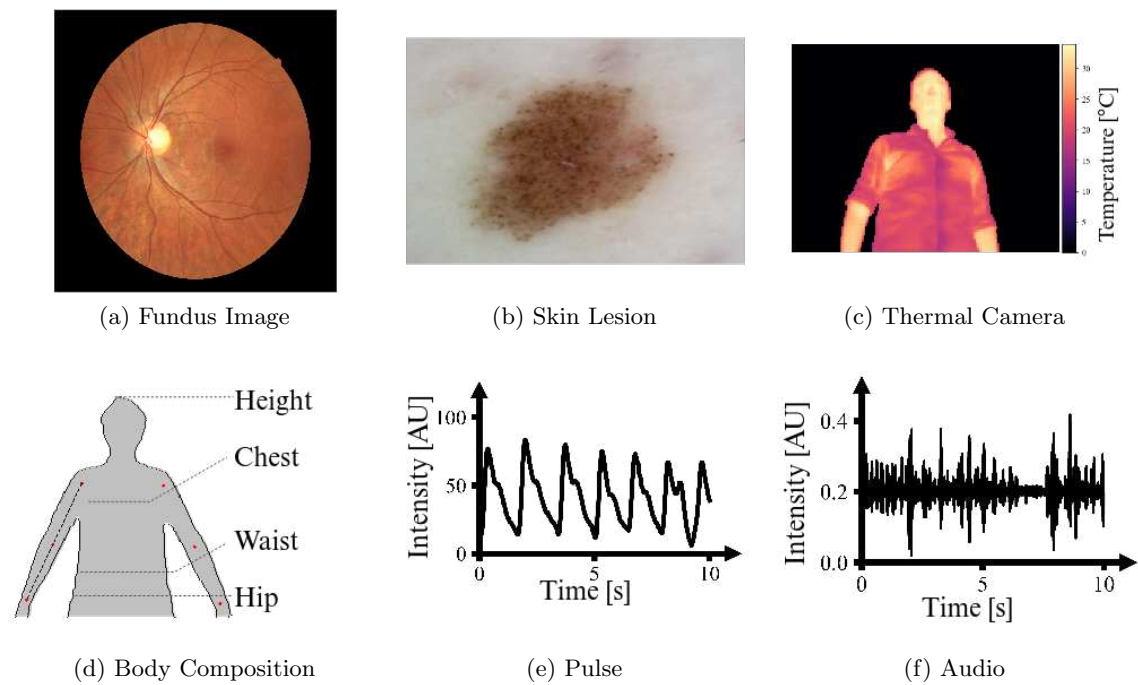


Figure 3.1.: The image illustrates the different biomedical recordings captured within the Virtual Doctor cabin, showcasing the integration of various sensors and imaging modalities for comprehensive patient assessment. The image was created with Biorender.com.

Additional custom hardware fixtures, such as sensor holders for the strain gauge, camera mounts, and protective casings for the thermal imaging device, were designed using FreeCAD and fabricated through 3D printing. This approach was chosen due to its high degree of customizability and cost-efficiency, which allowed for the rapid prototyping of specialized components tailored to the cabin's design constraints. Electrical connections, such as the wiring between the thermal camera and its signal amplifier, as well as communication interfaces for the scale, were manually soldered onto the circuit boards. The Printed Circuit

Board (PCB) layouts designed with EAGLE for both the thermal camera connection to the Raspberry Pi 4b and the scale's connection to the ESP32 microcontroller are provided in the appendix C.2.

The camera systems integrated within the Virtual Doctor were carefully calibrated to account for rotational distortions and tangential lens aberrations, following the mathematical formalism described in Section 2.2.1. Chess grid pattern recordings as shown in figure 2.6 were used to fit the camera's intrinsic, extrinsic, and distortion parameters. To obtain real-world coordinate mappings, the transformation process was based on an approximation of the distance between the patient and the camera system, with the reference point being the known distance between the camera modules and the embedded weight platform.

3.1.2. Distributed Systems Architecture

The Virtual Doctor cabin employs a distributed hardware architecture in which multiple sensing and processing units communicate with a central local server. The server is implemented on a Raspberry Pi 4B, which acts as the primary coordination node responsible for data aggregation, preliminary preprocessing, and communication with higher-level diagnostic algorithms.

All communication between the hardware components and the central server occurs over a secured local area network, utilizing Wifi Protected Access 2 (WPA2) encryption to ensure data confidentiality and integrity. This security layer mitigates potential risks of data interception or unauthorized access, which is particularly critical in a medical diagnostic environment where sensitive patient information is transmitted.

The integrated weighing scale is equipped with an ESP32 microcontroller, selected for its energy efficiency, integrated Wi-Fi functionality, and seamless compatibility with embedded sensor systems. The ESP32 is configured to connect to the local network, where it continuously records weight measurements from the strain gauge sensor. Data are transmitted to the central Raspberry Pi server through socket-based communication, ensuring real-time transfer with minimal latency. A comparable architecture is employed for the RGB camera, which is operated via a Raspberry Pi Zero W. This module captures high-resolution silhouettes of the patient, with the processed outlines subsequently stored in the system to preserve privacy. Additional diagnostic devices, including the dermatoscope and pulse oximeter, are integrated through a USB interface, thereby enabling standardized and reliable communication across the cabin's data acquisition framework.

The distributed architecture allows for modular expansion, meaning that additional sensing devices—such as new biomedical sensors or advanced imaging modules can be integrated without major modifications to the existing system. Each device can independently transmit its data streams to the central server, where they can be synchronized and fused for joint

analysis. This modularity is essential for future-proofing the Virtual Doctor platform and accommodating evolving diagnostic requirements.

Moreover, the use of lightweight, embedded computing nodes at the sensor level offloads preprocessing tasks from the central server, improving scalability and reducing computational bottlenecks. For example, preliminary filtering or noise reduction can be applied directly on the ESP32 before data transmission, reducing the amount of raw data that must be processed centrally.

3.1.3. Database Structure

The Virtual Doctor cabin incorporates a local web server interface to facilitate both system control and diagnostic workflow management. This interface is implemented in PHP and enables authorized users to operate various hardware components remotely within the cabin environment. Through the web server, it is possible to activate or deactivate the custom-integrated ventilation fan, control the lighting system, monitor real-time temperature and humidity levels, and initiate a diagnostic session. This integrated control layer enhances operational convenience and ensures that environmental conditions can be optimized prior to patient interaction.

The backend data management is handled through an SQLite database, interfaced with Python programs for data access, manipulation, and integration with the diagnostic pipeline. SQLite was selected for its lightweight footprint, reliability, and compatibility with embedded systems, making it well-suited for deployment on the Raspberry Pi 4B. The database stores the patient metadata and hardware operation logs.

Figure 3.2 presents the Entity-Relationship (ER) diagram [113], illustrating the logical data model and the relationships between database entities. The ER diagram provides a conceptual overview of how various data components, such as patient records, environmental parameters, and diagnostic results, are interconnected. This structure enables efficient data retrieval, supports transactional integrity, and facilitates seamless integration with the machine learning inference modules for decision support.

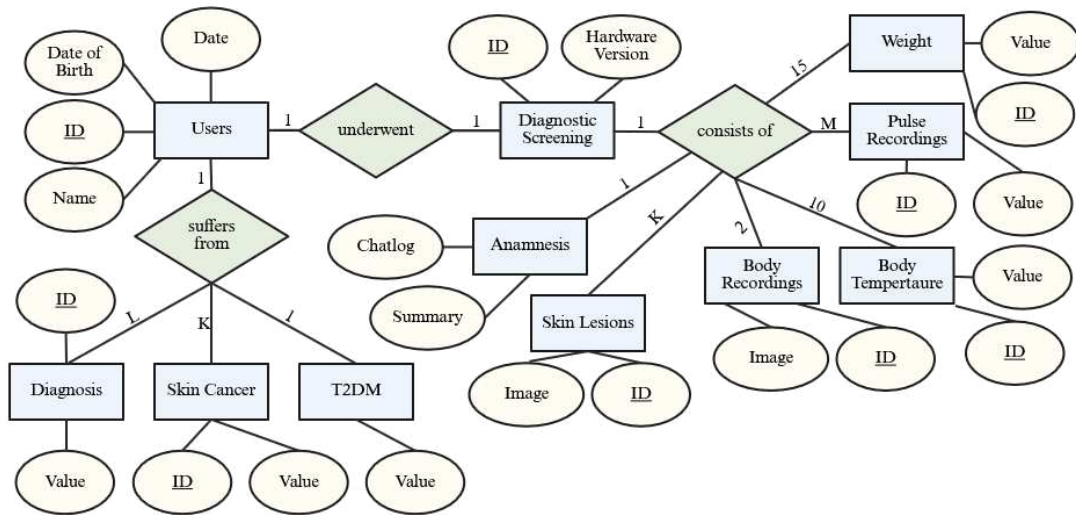


Figure 3.2.: The image provides the ER diagram for the Virtual Doctor Database. The image was created with BioRender.com.

Furthermore, the database schema is designed to support temporal queries and historical data analysis, allowing physicians or researchers to review past diagnostic sessions and observe trends over time. This capability is particularly valuable for longitudinal monitoring of patient health metrics or for evaluating the performance of diagnostic algorithms under different environmental conditions.

3.1.4. Software Interface

The software architecture of the Virtual Doctor cabin integrates multiple programming languages and frameworks, each selected to address specific functional requirements within the system. Development was modularized according to hardware control, data processing, user interface design, and inter-device communication, ensuring maintainability and scalability.

For low-level microcontroller programming, the C programming language was employed due to its efficiency, direct hardware access, and suitability for embedded systems with constrained computational resources. This layer manages sensor data acquisition, actuator control, and initial signal preprocessing before transmitting information to the central computing unit.

Python served as the primary language for higher-level operations, including data parsing, database interaction, diagnostic pipeline execution, and communication between distributed components. Python's extensive ecosystem of libraries—particularly for socket communication, machine learning inference, and sensor integration—enabled rapid prototyping and robust implementation.

The web-based control and monitoring interface was implemented using PHP for server-side

logic, in conjunction with HTML and CSS for static content rendering and layout design. This interface allows operators to remotely configure cabin hardware, monitor sensor readings in real time, and initiate diagnostic procedures.

Version tracking of all software components was maintained to ensure reproducibility and simplify future maintenance or upgrades. Table C.1 in the appendix C.1 provides a comprehensive overview of the software stack, including version numbers for all major modules, libraries, and development tools employed. This structured documentation supports transparency in system design and facilitates integration with future enhancements, such as more advanced machine learning decision-support modules or remote maintenance capabilities.

3.2. Decision Support using Deep Learning

During the initial development phase of the Virtual Doctor diagnostic cabin, decision-support capabilities were centered on three primary medical domains: T2DM, skin cancer, and ocular disease detection. These domains were selected based on their clinical relevance, prevalence in the general population, and the technical feasibility of integrating the required sensing and imaging equipment within the physical constraints of the cabin.

For T2DM risk assessment, the system incorporated non-invasive sensors capable of capturing biometric and physiological parameters, including weight, height, alongside self-reported lifestyle and medical history data. These inputs were processed using deep learning models trained to estimate an individual's likelihood of developing T2DM, serving as an early risk stratification tool.

Skin cancer diagnostics were enabled through a dermatoscopic camera mounted within the cabin. CNN architectures were trained on annotated dermatoscopic image datasets to differentiate between benign and malignant lesions. The models were optimized for inference speed to ensure real-time analysis and feedback within the consultation process.

For the detection of ocular diseases, deep learning models were developed and applied to identify pathological indicators associated with conditions such as diabetic retinopathy and cataracts. These models were trained on annotated ophthalmological datasets and designed to detect subtle visual patterns indicative of disease. To enhance interpretability, the raw model outputs were transformed into visual overlays that highlighted regions of interest within the retinal or anterior eye images, thereby supporting physicians in their diagnostic decision-making. Although a fundus camera is available within the broader Virtual Doctor project infrastructure, its deployment within the walk-in cabin was not realized due to prevailing insurance and regulatory constraints.

The final component of the AI-driven decision-support framework is the integration of an automated and targeted anamnesis process, which culminates in the generation of a

structured summary for clinical use. To achieve this, a LLM was fine-tuned on domain-specific medical datasets to capture the nuances of patient history-taking and to ensure medically relevant phrasing. The model was designed to guide the anamnesis by posing targeted, context-aware questions based on the patient’s previously provided information and recorded measurements. Importantly, the system does not provide a diagnostic outcome to the patient but instead compiles a comprehensive anamnesis report. This summary is stored securely in the local database and made available to authorized physicians, thereby assisting clinical decision-making while preserving patient privacy.

All deep learning models were containerized and deployed locally on the cabin’s central computing unit to ensure operational independence from external sources, thereby enabling use in remote or resource-limited settings. Model inference results were integrated into the cabin’s software interface and stored within the backend database, ensuring seamless access for physicians via the decision-support dashboard.

3.2.1. Normalization Strategies

During the development of the T2DM decision-support system, it was observed that the integration of adaptive normalization (learnable normalization) substantially improved the performance of the implemented MLP architecture. To systematically investigate this phenomenon, we conducted experiments across multiple medical and benchmark datasets, assessing the impact of introducing learnable normalization into the training pipeline.

Adaptive normalization departs from conventional normalization techniques by replacing fixed, analytically derived parameters, such as mean and variance, typically computed from the dataset, with parameters that are instead learned during the training process. This design enables the normalization layer to dynamically adjust to the statistical properties of the data as the model learns, rather than relying on static estimates computed a priori. Consequently, adaptive normalization introduces an increased degree of flexibility compared to conventional methods.

Since each sublayer of a neural network produces feature representations that serve as the input for subsequent layers, the research investigated the incorporation of adaptive normalization at multiple stages within the architecture. Specifically, adaptive normalization was examined both as a potential replacement for conventional normalization schemes and as an additional transformation layer augmenting existing structures. The rationale for this approach was to assess how adaptive normalization influences the distribution of intermediate features and whether it can mitigate training-induced shifts in statistical moments, which often hinder stable convergence. By enabling normalization parameters to adjust dynamically during training, the method was expected to improve the consistency of feature propagation across layers and enhance overall network robustness.

The investigation involved implementing adaptive normalization across different network

configurations and comparing it conceptually to established normalization techniques (see section 2.2.2.4). The design of these experiments was intended to assess how learnable normalization strategies can be incorporated into various neural network architectures and data modalities without being constrained to fixed statistical assumptions.

3.2.2. Pathological Anomalies in Fundus Images

For the development of a neural network-based decision-support system aimed at detecting pathological anomalies in retinal fundus images, a dedicated dataset was compiled through an in-house study conducted at the Heinrich Heine University Düsseldorf. This dataset was collected under controlled conditions and subsequently combined with publicly available fundus image datasets. The inclusion of multiple data sources was intended to account for potential domain shifts, such as variations in imaging devices, acquisition protocols, and image quality, which can influence the generalizability of machine learning models.

The methodological workflow involved standardized image preprocessing, data harmonization, and dataset integration to minimize bias introduced by heterogeneous data origins. Special attention was given to ensuring consistency in resolution, color normalization, and field-of-view representation before model training.

Although a fundus camera is currently available within the Virtual Doctor project infrastructure, integration into the physical diagnostic cabin is not planned at this stage due to insurance and regulatory constraints. Nonetheless, the collected fundus datasets remain part of the broader research framework for evaluating and refining pathological anomaly detection algorithms.

3.3. The Virtual Doctor Avatar

To enhance the interaction within the Virtual Doctor framework and provide a user-friendly interface for anamnesis, a virtual doctor avatar was developed. This avatar is realized through a combination of multiple transformer-based models, which were trained and fine-tuned specifically for the context of automated medical interviews. The system architecture includes a speech-to-text pipeline, allowing real-time conversion of the patient's spoken input into textual data, and a text-to-speech pipeline, enabling natural verbal responses from the avatar.

In addition, a text-to-video extension was implemented to synchronize the generated speech with a visual representation of the avatar, facilitating a more immersive and interactive patient experience. This visual representation allows the avatar to provide gestures and facial expressions, improving patient engagement during anamnesis while maintaining a professional and empathetic demeanor.

The automated and targeted anamnesis process is produced by a fine-tuned LLM, which interprets the patient’s medical history and generates contextually relevant questions. Once the anamnesis is completed, a summary of the patient’s responses is stored securely in the database, making it accessible to the corresponding physicians for review and follow-up.

The integration of the Virtual Doctor avatar within the framework ensures seamless interaction between the patient, the diagnostic sensors, and the backend decision-support systems. This modular design allows for flexibility in future updates, such as incorporating additional languages, expanding the scope of medical conditions covered, or integrating more advanced multimodal inputs such as eye-tracking or gesture recognition.

Overall, the Virtual Doctor avatar serves as a bridge between patients and healthcare providers, enabling structured data acquisition while maintaining patient engagement, empathy, and trust.

3.4. Ethical Approvals

Although not explicitly included in the presentation of research results, both the fundus image study and the Virtual Doctor project required the preparation of comprehensive ethical documentation. For the fundus image study, all necessary protocols were developed, submitted to the appropriate ethics committees, and subsequently revised in response to feedback. Similarly, for the Virtual Doctor project, detailed documentation addressing ethical considerations was drafted in preparation for formal submission. At the current stage, formal approval for the Virtual Doctor project has not yet been granted, but the groundwork for ethical compliance has already been established to facilitate future approval processes.

The preparation of these study documents required careful planning and thorough consideration of data security, patient privacy, and regulatory compliance. To ensure compliance with ethical standards, the Virtual Doctor framework incorporates several measures which include that all patient data are anonymized at the point of collection, and unique identifiers are used instead of personally identifiable information. Access to sensitive data is strictly restricted, with role-based authentication controlling which users can view or modify data. Communication between hardware components, the local server, and the database is encrypted using WPA2-secured networks and secure protocols.

Additionally, all diagnostic and decision-support outputs generated by the system are stored in encrypted form. Logs of system access and data processing are maintained to allow auditability and traceability of actions within the system. The system design also follows principles of data minimization, ensuring that only data necessary for the clinical and research purposes are collected and stored.

Through these measures, the Virtual Doctor framework is structured to meet high ethical

standards for medical research and clinical use, addressing concerns regarding patient confidentiality, data integrity, and regulatory compliance, even prior to formal ethical approval.

4. Results

This chapter is structured according to the four principal development stages of the Virtual Doctor project, with each stage corresponding to a distinct set of research objectives and technical milestones. The results obtained at each stage are presented in dedicated sections, each of which aligns with an associated scientific publication that has been published, is currently under review, or has been submitted to a peer-reviewed journal. This structure ensures a coherent presentation of the progressive advancements made throughout the project lifecycle.

4.1. Virtual Doctor Hardware and Software Framework

The design and assembly of the Virtual Doctor walk-in cabin constituted the central element of this thesis. Significant effort was devoted to the conception, development, and realization of the hardware infrastructure, ensuring seamless integration with the accompanying software framework. The process was iterative and characterized by numerous failed prototypes and unsuccessful trials in both hardware and software development. While many of these intermediate steps remain undocumented in the final version of the system, they represent essential stages in the learning and refinement process that ultimately led to the functional and operational Virtual Doctor walk-in cabin.

A key methodological innovation in the decision-support component was the incorporation of the adaptive normalization layer, particularly in the context of T2DM risk prediction. Its integration resulted in a measurable and statistically significant improvement in model performance. Using the Wilcoxon signed-rank test on a 10-fold cross-validation with ROC-AUC values as the evaluation metric, a p-value of 0.004 was obtained, thereby underlining the robustness of the observed performance gain.

In addition, calibration of predictive outputs was addressed through the integration of the GUESS framework. To this end, a dedicated Python implementation of GUESS was developed, which expanded upon the original version by supporting a broader range of PDF fits. This enhanced implementation ensures greater flexibility in modeling score distributions and is made publicly available, thereby contributing to reproducibility and enabling its use in related projects.

Beyond the technical components, the Virtual Doctor project also incorporated an evaluation of socio-psychological factors associated with its use. This dimension was essential

to understanding user acceptance, perceived trustworthiness, and the intention to engage with such a diagnostic system, ultimately contributing to a more patient-centered and socially aware system.

4.1.1. Contributions

My individual contribution to the project was extensive and multifaceted. I was primarily responsible for the design, development, and integration of both the hardware and software components, including the management of project materials, logistics, and the evaluation of appropriate sensor technologies. The overarching framework of the Virtual Doctor was designed by me, with targeted support during the physical realization of hardware elements. With respect to socio-psychological studies, my contribution focused on ensuring the seamless integration of findings into the overall manuscript, thereby producing a coherent and scientifically rigorous document. Furthermore, I was the main author of the manuscript, responsible for the majority of the written work and for aligning the diverse project components into a unified scholarly presentation.

The Principal Investigator of the project was Prof. Dr. Dominik Heider, who confirms that the description of my contributions is accurate.

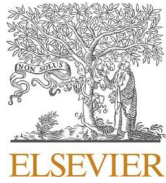
Journal:	Computers in Biology and Medicine
Contribution to the publication:	70%
Type of authorship:	First Author
Status of publication:	Published

Münster, 27.11.2025

Place, Date



Prof. Dr. Dominik Heider



The virtual doctor prescribing the future: Diagnostics with interactive clinical decision support

Jan Benedikt Ruhland^a, Johannes Wichmann^b, Dmitry Degtyar^c, Roman Martin^d, Leon Fehse^d, Jan Henric Klau^c, Thorsten Papenbrock^c, Jan-Peter Sowa^e, Ali Canbay^e, Bernd Freisleben^c, Michael Leyer^b, Dominik Heider^{c,d,*}

^a Heinrich Heine University Düsseldorf, Faculty of Mathematics and Natural Sciences, Düsseldorf, Germany

^b Philipps University of Marburg, Department of Economics, Marburg, Germany

^c Philipps University of Marburg, Department of Mathematics and Computer Science, Marburg, Germany

^d University of Münster, Institute of Medical Informatics, Münster, Germany

^e University Hospital Knappschafts Krankenhaus Bochum, Department of Internal Medicine, Bochum, Germany

ARTICLE INFO

Keywords:

AI in medicine
Clinical decision support system
Deep learning
Skin cancer
Diabetes

ABSTRACT

The growing shortage of healthcare professionals, particularly in rural areas, poses a significant challenge to timely and effective medical care. To address this issue, we present the Virtual Doctor, a walk-in cabin designed to collect and analyze medical data without direct medical supervision. This system integrates non-invasive biomedical sensors for continuous data acquisition and leverages artificial intelligence based decision support to facilitate early risk detection and enhance healthcare accessibility.

The Virtual Doctor employs a distributed medical data acquisition system composed of multiple components that are detailed in this study. The collected biomedical data is processed by two pre-trained deep learning models developed on external datasets to predict the risk of type 2 diabetes mellitus and skin cancer. The diabetes model achieved an ROC-AUC of 0.729 on the test set, while the skin cancer model attained a mean ROC-AUC of 0.81. Beyond technical development we evaluated the socio-psychological acceptance of the Virtual Doctor through a survey of 1217 participants focused on two use cases, routine skin cancer screenings and antibiotic prescriptions. The developed model explaining users' intention to use the Virtual Doctor showed meaningful results for both scenarios ($R^2 \geq 0.604$). The results suggest broad general acceptance of the system while also highlighting the need for comprehensive explanations during serious disease screenings to reduce anxiety and potential resistance.

1. Introduction

The COVID-19 pandemic has exposed the vulnerabilities of health systems globally and has highlighted the significant problems physicians had to deal with. Many healthcare systems worldwide were already stressed by demographic and geographic factors, including the challenges of increased life expectancy and thus a higher number of total medical treatments, shortages of physicians in rural areas, and the age distribution of the medical staff.

In September 2022, the World Health Organization (WHO) called upon all European countries to develop and adjust countermeasures in response to the existing state of health systems [1]. This initiative led to the "Bucharest declaration on health and care workforce" [2], published

in March 2023 and endorsed by 50 out of 53 member states of the WHO European Region. In this declaration, member states reiterated their commitment to improving current working conditions and to using technological interventions. Furthermore, the WHO projected a shortage of 12.9 million healthcare workers worldwide by 2035, underscoring the severity of this global predicament [3]. Thus, healthcare technologies are needed, especially vital in rural regions, to overcome the shortcomings in healthcare personnel [4–8].

Artificial intelligence (AI) has become important in medicine, offering innovative solutions to improve diagnostic accuracy and streamline the overall healthcare workflow, helping to reduce the number of healthcare personnel to conduct treatments [9,10]. In particular machine learning (ML) algorithms, as a subclass of AI, are

* Corresponding author. Philipps University of Marburg, Department of Mathematics and Computer Science, Marburg, Germany.
E-mail address: dominik.heider@uni-muenster.de (D. Heider).

extensively applied to analyze biomarkers and medical images, providing valuable insights for diagnostic purposes [11,12]. Additionally, technological advancements play a crucial role in reducing the workload of healthcare professionals by automating patient data monitoring. For instance, the widespread adoption of Smart Healthcare Monitoring (SHM) devices that utilize passive patient surveillance, marks a significant breakthrough [13]. However, many SHM devices require self-purchase and often raise concerns regarding data protection and privacy standards.

An alternative approach involves leveraging telemedicine cabins as clinical decision-support systems powered by AI-driven health information systems. While this represents an innovative application, the concept of telemedicine cabins itself is not new. They have been investigated in various environments by countries, such as in Scotland [14], France [15,16], China [17], USA (e.g., TalkBoxBoothLL), Singapore (e.g., MinmedGroupPteLtd), or India [18]. Consequently, telemedicine cabins are used for several medical treatments, such as infectious diseases or allergies [16]. Nonetheless and until now, treatments were performed by real physicians, who either took measures by themselves [14] or guided patients virtually to take measures appropriately [16].

In this study, we describe the design of the Virtual Doctor walk-in cabin along with its corresponding software framework. This study presents a walk-in cabin that uses AI for autonomous anamnesis, integrating biomedical sensors with AI components to showcase the innovation of this approach. Additionally, the study includes a crucial prestudy on the social acceptance of this concept. This serves as a reference for future projects aiming to utilize data acquired from the Virtual Doctor or implement similar concepts to develop an unsupervised data acquisition system with diagnostic assistance. Currently, the Virtual Doctor covers two use case scenarios: type 2 diabetes mellitus and skin cancer detection. We trained and evaluated ML models for each specific prediction task. Additionally, we commissioned the crowdsourcing platform Clickworker to conduct a survey to analyze the intention to use the Virtual Doctor based on economic and psychological aspects. The survey results were evaluated using an Integrated Theoretical Model (ITM) derived from healthcare, psychology, and information system theories to determine the impact of maladaptive and adaptive appraisals regarding the use of the Virtual Doctor [19].

This manuscript is structured as follows: In Section 2, we provide our materials and methods by introducing the design of the Virtual Doctor walk-in cabin and presenting how we measured the acceptance of AI in medicine. Section 3 presents our results for diabetes and skin lesion prediction as well as for our integrated theoretical model. In Section 4, we discuss our results. Section 5 concludes the study and outlines areas for future research.

1.1. Research in context

The advantages of telemedicine, autonomous patient monitoring, and AI-driven decision-support systems are numerous [13,18,20]. On the one hand, telemedicine offers economic benefits by reducing the time required compared to traditional doctor visits while also minimizing the risk of exposure to highly contagious diseases. However, its primary limitation lies in the restricted use of biomedical sensors that are essential for a comprehensive medical assessment [21].

Technical barriers also hinder the full deployment of telemedicine. For example, a study conducted by Wang et al. [22] highlighted key challenges in network design that must be addressed to enable effective telemedicine implementation. Once deployed, however, telemedicine has shown high levels of social acceptance and demonstrated effectiveness for both general evaluations and targeted applications, such as autism diagnostics, as shown in the study by Wagner et al. [23]. That study emphasized the need to improve access to healthcare and proposed combining telemedicine with machine learning tools as a viable solution. Another application area is ophthalmology, where remote anamnesis has been explored as a diagnostic aid [24].

Autonomous patient monitoring, on the other hand, is typically employed in the absence of medical personnel and is often specialized in particular fields, such as cardiovascular diagnostics [25]. Smart patient monitoring systems frequently integrate unsupervised medical data collection with advanced algorithms or AI-driven analysis [13]. Recent innovations, such as developments in biochip technology, further enhance the capabilities of medical monitoring systems by providing more detailed insights into physiological states [26].

Telemedicine solutions are frequently combined with smart healthcare devices [13], and a shared characteristic across these advancements is the integration of AI-based algorithms. However, due to current regulatory and legal frameworks, fully autonomous diagnostic systems are rarely permitted. AI tools are generally required to operate under physician oversight, which limits the feasibility of deploying a fully autonomous system.

In recent years, significant progress has been made in medical large language models (LLMs), such as MedChatZH [27], Med-PaLM 2 [28], and AMIE [29], which demonstrate strong capabilities in clinical reasoning, guideline-based recommendations, and interactive medical question answering. These models are shaping the next generation of conversational AI in healthcare and may play an important role in future iterations of autonomous diagnostic systems like the Virtual Doctor [30].

Simultaneously, advancements in Explainable AI (XAI) have introduced techniques such as SHAP [31], Integrated Gradients [32], and Grad-CAM [33], which are increasingly applied in medical contexts to enhance model transparency and clinician trust. Recent examples include gene contribution analysis for breast cancer development [34] and brain tumor region segmentation [35].

In parallel, pod-based diagnostic devices and AI-enabled health kiosks, such as the HERMES Kiosk [36], have gained attention as scalable, modular solutions for decentralized care. Commercial and research efforts in this area primarily focus on vital sign screening or triage support. The Virtual Doctor expands this paradigm by integrating multiple non-invasive modalities into a low-cost, autonomous cabin designed for offline use, enabling broader applicability in low-resource settings. The Virtual Doctor project [30], initiated in 2019, represents an early attempt to create an autonomous clinical decision support system. While it facilitates an AI-driven anamnesis process, the system's outputs are still reviewed by medical professionals, with the final diagnostic decisions remaining the responsibility of the attending physician.

2. Materials and methods

2.1. Project overview

To provide a more thorough understanding of our research methodology, we have systematically presented the complete set of capabilities of the Virtual Doctor in Table 1. Unlike other projects that primarily focus on remote communication between patients and healthcare professionals [14,15] or those that rely solely on AI-based decision-support after a medical examination [37], our approach sets itself apart by incorporating pre-diagnostic assessment and decision-support from the beginning. This allows the system to provide healthcare professionals with valuable insights and recommendations before further medical examinations are conducted. By offering this early guidance, our system enhances the decision-making process and improves the overall efficiency of patient care.

2.2. Design

2.2.1. Design of the Virtual Doctor cabin

The Virtual Doctor functions as a comprehensive clinical decision-support system. An initial prototype of the Virtual Doctor was introduced in Ref. [30] and has been completely redesigned since its inception. A cabin was developed and built to serve as the

Table 1

Overview of the hardware and software components included in the Virtual Doctor cabin.

Decision Support	Hardware	Software	Biomarkers
Pyrexia	Thermal Camera	Keypoint detection with Mediapipe (deep learning)	Face surface temperature
Hypoxemia	Beurer PO80, USB interface	Python interface	Blood oxygen concentration
Adipositas	Strain Gauge, Cameras, Wireless connectivity via ESP32 microcontroller	Body measurement inference through calibration and keypoint detection	Body weight, height, taille
Mental well-being	Microphone	Speech-to-text transformer (deep learning)	Age, biological sex, well-being biomarkers
T2DM	Strain Gauge, Cameras, Microphone	Custom multi-layer perceptron	Body weight, height, taille, smoking and drinking habits, biological sex, age
Skin Cancer	Dermatoscope, USB interface	Custom trained DenseNet201	Skin lesion recording

human-machine-interface to conduct diagnostic procedures. We used the open-source CAD-software FreeCAD 0.21.2 to create a 3D model of the cabin, allowing for the customization of designs to accommodate specific hardware components, such as the incorporation of a weight scale into the floor. Fig. 1 provides a schematic overview of the Virtual Doctor walk-in cabin, showcasing its design and intended purpose.

2.2.2. User interface

After the physician registers the user in the database and the user enters the Virtual Doctor walk-in cabin, they are first informed that participation in the examination is entirely voluntary and that they may exit at any time. Additionally, the user is provided with instructions on special voice commands such as “help,” “stop,” or “skip” to facilitate a smoother examination process. Alternatively, the user may use the numeric keypad located in the cabin. To enable voice control, we

implement speech-to-text functionality using the transformer-based architecture from Huggingsound [38]. The audio input is captured using a BOYA BY-MC2 Conference Microphone and processed through the model to recognize key commands. These detected keywords allow the system to interpret user intent and adjust the examination accordingly. The interaction between the Virtual Doctor and the user is designed to be intuitive and accessible. The Virtual Doctor provides verbal explanations of examination procedures, complemented by written instructions displayed on the screen. Users can respond verbally or input their answers using a numeric keypad. Exemplary screens shown during the Virtual Doctor examination are presented in Appendix A. Speech-to-text conversion is facilitated by a wav2vec 2.0 implementation, enabling seamless transcription of verbal responses in both German and English [39]. For this purpose, we employ two pre-trained models from Huggingsound, which can be deployed locally [38].

The graphical user interface of the Virtual Doctor is displayed on a Samsung 43" Crystal UHD 4K BU8079 screen, providing real-time visual feedback based on the user's interactions. On the back-end, deep learning pipelines are executed using an Intel NUC 12, ensuring efficient processing of speech recognition and other AI-driven tasks within the system.

2.2.3. Sensors and data

Initially, the user is instructed to stand on the designated platform to measure their body weight. The platform consists of four strain gauges connected to an ESP32 microcontroller. Once the weight data is recorded, it is transmitted via the local network to the central computing unit for further processing. Following the completion of this step, the system proceeds to the next examination.

For the body measurement evaluation, the user is asked to stand with their arms extended away from the body to ensure accurate assessment. An image is then captured, after which the user is instructed to rotate 90° to allow for a second image to be taken from a different angle. These images are analyzed using key point detection techniques to extract relevant body measurements such as height, waist and taille circumference. To ensure privacy and data security, only the computed measurements are stored, while the original images are immediately deleted after processing. We employ one Raspberry Pi V2.0 and one Raspberry Pi V3.0 camera, each linked to a Raspberry Pi Zero, transmitting images via

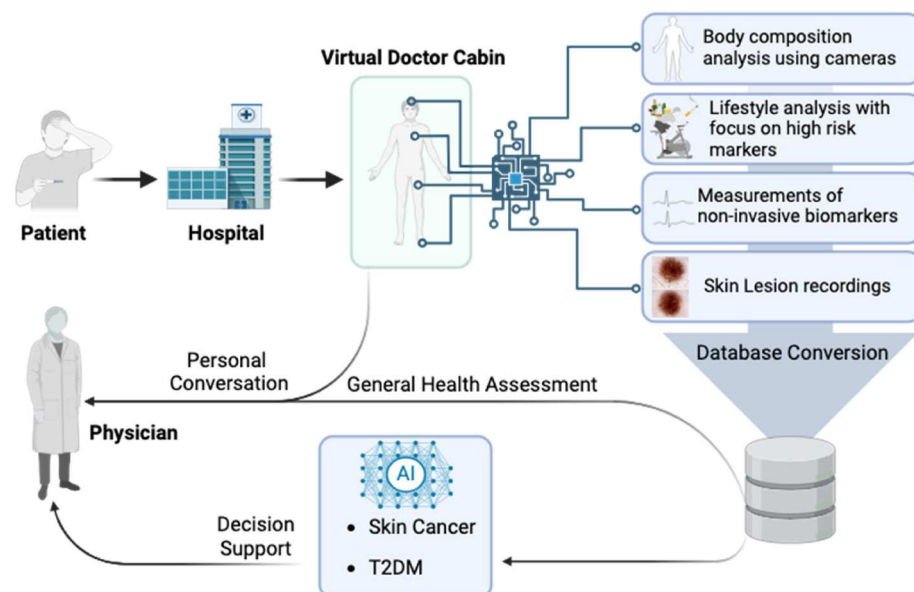


Fig. 1. Overview of the design and concept of the Virtual Doctor. The hardware design commences with the cabin layout and is tailored to accommodate all listed applications with corresponding sensor modules. For data acquisition, we employ a pre-recorded procedure that systematically guides users through all optional diagnostic processes. The results are stored in a database on the central computing unit. Physicians can access the local network and review the acquired biomarkers and diagnostic evaluation. The figure was created in BioRender.com.

a local network to the central computing system (Raspberry Pi 4b). The process of transforming camera coordinates into real-world coordinates necessitates the use of mathematical approximations and a well-defined theoretical framework. As a first step, the camera coordinate system was established within the cabin to enable the transformation of any given point within the cabin into the camera frame of reference. Given an approximate estimation of the camera positions, we employed a Poincaré transformation to map points from the real-world coordinate system to the camera frame. The marked platform the user stands on reduces the projection problem onto a two-dimensional hyperplane. To ensure the accuracy of the captured image data, the cameras underwent a distortion correction process. Specifically, radial and tangential distortions were corrected using Brown's distortion model, which compensates for optical aberrations introduced by the camera lens [40–43]. We took multiple photos of a 2D chess board to fit the model to the data. Furthermore, the spatial resolution of the imaging system was approximated by determining the distance between two adjacent pixels. This was achieved by considering the physical sensor spacing of the camera in conjunction with its focal length, thereby providing an estimate of the real-world scale of the captured image. With this approach, we can estimate the systematic uncertainty from the spatial distribution of the sensor as the transformed distance between two pixels in real-world coordinates. To enhance imaging accuracy, we equipped one camera with a linear polarization filter and the light source with a 90° shifted polarization filter within the cabin. Since surface-reflected light primarily retains its polarization, it is effectively filtered by the applied polarization filter [44]. A similar principle is employed in dermatoscopy to capture light reflected from deeper tissue rather than the skin surface.

In the next step of the examination, the user is instructed to maintain an upright position and look directly at the screen. A Lepton 3 thermal camera, connected to a Raspberry Pi 4B, records the surface temperature of the user's face. This specific type of thermal camera has previously been employed in similar studies, where it demonstrated a high level of accuracy, achieving 98.1 % precision in temperature readings within the range of 36.1 °C–37.5 °C [45,46]. Following the temperature recording, the user is instructed to place their left index finger into the Beurer PO80 pulse oximeter. The device continuously measures the user's pulse rate and oxygen saturation (SpO₂) levels for a duration of 1 min and transmits the recorded data to the central computing unit via a USB interface. After completing the physiological measurements, the user is presented with a generalized questionnaire designed to assess their overall well-being. The details of the questionnaire are provided in Section 2.3. The user has the option to respond using either voice commands or the keypad, ensuring accessibility and ease of interaction.

At the conclusion of the examination, the user is given the opportunity to record suspicious skin lesions using a dermatoscope. The captured images are projected onto the screen in real time, allowing the user to review the recordings. The recorded data is then securely stored in an SQLite database and made accessible to authorized physicians via the local network. This structured and autonomous data acquisition system enables the standardized collection of non-invasive biomarkers, contributing to improved diagnostic efficiency while reducing the workload for medical professionals. Further details regarding the developed web interface are provided in Appendix A.

2.3. Generalized questionnaire for health assessment

The questionnaire framework is based on three established health assessment tools: the General Health Rating Index (GHRI), the Quality of Well-being Scale (QWB), and the Sickness Impact Profile (SIP) [47,48]. Additionally, the design of the questionnaire incorporates considerations of the five components of social well-being [49] and the six components of eudaimonic well-being [50].

To optimize efficiency and minimize user time spent in the cabin, the questionnaire was condensed to the 18 questions listed in Table 2. Questions 2, 6, 7, and 15 require numeric responses, while all others,

Table 2

Questions asked by the Virtual Doctor during the examination. The questions were derived from the GHRI, QWB, SIP, the five components of social well-being as well as the six components of eudaimonic well-being.

Number	Question
1	What is your biological sex?
2	How old are you?
3	Do you get sick more often than the people around you?
4	When you get sick, how likely is it that you see a doctor?
5	Do you agree: you feel much better compared to your past self?
6	On average, on how many days per week do you drink alcohol?
7	On average, how many cigarettes do you smoke a day?
8	Do you agree: I feel really cared for by my spouse, partner or family.
9	Do you agree: I feel really cared for by a friend.
10	Do you agree: I have a clear vision of my life goals.
11	Do you agree: I find it helpful to set a goal for the near future.
12	Do you agree: I feel capable of changing bad situations for the better.
13	Do you agree: Difficult situations are also a learning experience.
14	Do you agree: I am fully satisfied with my financial status.
15	How many years did you spend within an educational institute?
16	Do you agree: The information provided by the Virtual Doctor is easy to understand.
17	Do you agree: It is easy to interact with the Virtual Doctor.
18	Do you agree: I am very satisfied with the Virtual Doctor.

except for question 1, use a 1 to 5 scale. The questions are presented sequentially, with validation checks in place to ensure responses conform to the expected format.

2.4. Clinical decision-support use cases of the Virtual Doctor

The Virtual Doctor system is designed to capture multiple biomarkers, providing a foundation for advanced data analysis through deep learning techniques. To enhance its diagnostic capabilities, we developed two specific use cases that leverage machine learning models for clinical decision support. For this purpose, we utilized external datasets, performed in-depth analyses, and integrated the trained models into the Virtual Doctor framework.

During the model development process, we carefully selected the most relevant features to ensure compatibility with the system's technical specifications. Due to hardware limitations and system constraints, we refined the feature set to optimize computational efficiency without compromising analytical accuracy. This strategic selection allowed the deep learning models to operate seamlessly within the Virtual Doctor environment while maintaining robust performance.

The trained models are designed to assist physicians by providing automated clinical decision support. Given their high prevalence in the general population, we focused on two critical medical conditions: T2DM and skin cancer. By analyzing relevant biomarker data, the models facilitate early detection and risk assessment, ultimately improving diagnostic workflows and supporting timely medical interventions.

2.4.1. Datasets

We used data from the Heinz-Nixdorf-Recall (HNR) study [51] to adjust for biomarkers obtained by the Virtual Doctor. The HNR study cohort was randomly selected from the cities of Bochum, Essen, and Mülheim in Germany, and includes 4814 participants. It provides over 100 invasive and non-invasive features aimed at predicting cardiovascular events. Among the participants, 656 were diagnosed with diabetes.

Our analysis focused on seven non-invasive features which were age, sex assigned at birth, weight, height, waist circumference, self-reported alcohol consumption, and smoking status. These variables have been identified in previous studies as significant risk markers of type 2 diabetes mellitus [52]. Moreover, the number of features measurable by the Virtual Doctor is limited due to the constraints of the system's design.

Specifically, the information about age, sex assigned at birth, alcohol consumption (measured on a scale from 1 to 6), and smoking status

(binary) is obtained through the implemented questionnaire. The user's weight is measured using the integrated scale, while height and waist circumference can be estimated from camera-based recordings within the Virtual Doctor framework.

After excluding samples with missing values, we randomly selected 10 % of the dataset (478 samples) as a held-out test set. The remaining 4305 samples were used for model training. Table 3 presents the demographic distribution of the preprocessed HNR dataset, including the corresponding mean values and relative proportions.

For the development of a skin lesion classification pipeline, we used the HAM10000 data set [53]. It contains 10,015 images that are classified into one of seven types of skin lesions. Among these, melanoma (mel), basal cell carcinoma (bcc), and actinic keratosis (akiec) are classified as malignant. The remaining categories, including benign keratosis like lesions (bkl), dermatofibroma (df), vascular lesions (vasc), and melanocytic nevus (nv), are considered benign. We removed a stratified subset that contained 20 % of the data and served as a validation set. The images used for testing were obtained from the ISIC-2018-challenge competition [54]. This dataset consists of 1511 images and has a similar format as the HAM10000 dataset that was used in the training.

2.4.2. Training

For the T2DM prediction task, we utilize a custom neural network composed of four DeepNetBlock modules. Each DeepNetBlock includes three linear layers, two adaptive normalization layers [55], and internal skip connections. If necessary, dimensionality adjustment is handled automatically within the block via a learnable linear transformation. Each DeepNetBlock is treated as a single hidden layer in the network. The architecture is preceded by an adaptive normalization step [55] and an initial linear projection with ReLU activation. After passing through the four residual blocks, the network concludes with a fully connected output layer followed by a sigmoid activation to produce a probability score for T2DM risk.

To address class imbalance, we employed Gaussian Noise Upsampling (GNUS) [56], which increased the number of diabetes (synthetic) cases to achieve a stratified training dataset. Notably, biological sex was excluded from this upsampling process. Hyperparameter optimization was performed using 10-fold cross-validation and grid search. We compared the baseline model to several modified versions, which included upsampling with different scaling factors and increased model depth. For GNUS, the scaling factors μ and σ were systematically explored in logarithmic steps ranging from 0.0001 to 0.1. For the comparison, we first assessed the mean AUC across models. To determine whether the observed differences were statistically significant, we applied the Wilcoxon signed-rank test to the cross-validation results. The models were trained using stochastic gradient descent with binary cross-entropy as the loss function. Training was conducted over 2000 epochs, and model performance was evaluated using the Area Under the Receiver Operating Curve (AUC). The learning rate was initially set to 0.03 and was reduced by a factor of 0.5 if the binary cross-entropy loss did not decrease within 50 consecutive epochs. Once the best hyperparameters were determined, the final model was trained on the full dataset and subsequently calibrated using GUESS [57], transforming the

Table 3

Demographic distribution of the HNR dataset used in this study after removing entries with missing values.

	Male (n = 2379)	Female (n = 2404)
Age [y]	59.65 ± 7.82	59.62 ± 7.80
Height [cm]	174.81 ± 6.77	162.06 ± 6.16
Waist Circumference [cm]	100.26 ± 10.73	88.46 ± 12.84
Weight [cm]	86.20 ± 13.18	72.56 ± 13.81
Smoking	1711 (72 %)	1063 (44 %)
Alcohol Consumption (scale 1 to 6)	3.41 ± 1.64	4.48 ± 2.07
Diabetes	415 (17 %)	237 (10 %)

neural network scores into probabilities of having T2DM using the Bayes theorem.

Due to significant imbalance in the HAM-10000 dataset, we implemented random upsampling from each class and augmentation techniques to mitigate the risk of overfitting of images. This process continued until each class contained the same number of images as the majority class. This technique, described by Pati et al. (2022), demonstrated improved results compared to not using upsampling [58]. The images were center-cropped and randomly rotated. The resulting images were normalized for each color channel to a uniform scale. We evaluated eleven pre-trained convolutional neural network (CNN) architectures using the upsampled training set and conducted on our high-performance cluster. Optimization was performed using stochastic gradient descent with a batch size of 128 and sigmoidal learning rate annealing, starting from an initial learning rate of 0.01 [59]. Because of the multiclass classification task, cross-entropy was employed as our loss function [60]. The training spanned 50 epochs, after which no significant improvement was observed. We then determined the optimal cut-off epoch based on the evaluation metrics on the validation set. After training, we assessed their performance on both the validation and test datasets, using metrics including accuracy, weighted F1 score, and Matthews Correlation Coefficient (MCC).

2.5. Socio-psychological aspects of the Virtual Doctor

To better understand the factors that lead to an intention to use the Virtual Doctor, we conducted a preliminary online survey in December 2022 in Germany with 382 individuals to initially test our acceptance questionnaire. After adjusting our acceptance questionnaire according to our insights of the preliminary online survey, we conducted our full survey with 1217 individuals in the US from June to August 2024. First, we introduced our participants to the concept of the Virtual Doctor by providing a description to them. Second, we collected demographic data that comprised gender, age, and educational qualification. Third, our individuals had to assess two test cases, i.e., antibiotics prescription and skin cancer check-ups, in which the individuals had to balance between treatments by physicians that are resource intensive and not resource-intensive, but resource-efficient treatments by the Virtual Doctor. Fourth, individuals had to pass attention checks before, during, and after the questionnaire to ensure that their answers are valid. Hence, we followed recommendations of Goodman et al. [61] concerning attention for unsupervised surveys. Fifth, by considering the test case scenarios and the ongoing necessity to pass attention checks, our participants had to solve our ITM questionnaire.

2.5.1. Data collection

Our total sample contains 1217 individuals, of which 72.6 % (n = 884) were female and 27.1 % (n = 330) were male. The mean age was 38.2 years (standard deviation (SD): 10.1), with age ranging from 18 to 98 years. For the classification of the educational level of our participants, we used the 201 ISCED categories [62], where 1.8 % (n = 22) of our participants had reached the highest level (level 8; Ph.D. or equivalent) and 1.7 % (n = 21) had reached level 2 (completion of lower secondary education). The most common level achieved (38.0 %, n = 462) was level 6, as a Bachelor's degree. We used the crowdsourcing platform Clickworker (similar to Amazon MTurk) to collect our data and paid individuals to participate in the survey, following the recommendations of Goodman et al. [61] to ensure data validity.

2.5.2. Evaluation methods

For the evaluation and identification of psychological factors behind patients' use of the Virtual Doctor as a specific information system, our research combines theories from healthcare, psychology, and information systems. For this purpose, we use the ITM of Gesk et al. [63] that has been developed for such purposes and tested with AI applications. The integrated model is superior to theories from the single disciplines as it

allows to consider technological aspects in the healthcare domain and with this allows for a better explanatory power of intentions as one of the main drivers for behavior [63]. From the model we derived 27 hypotheses on how the different aspects for the intention to use the Virtual Doctor affect each other. The hypotheses are listed in Appendix D.1 and represented by the boxes on the edges between the main nodes in Fig. 2.

The analysis of the data was performed using SmartPLS 4.1.0.8. Because we consider two groups, using the Virtual Doctor for antibiotics prescriptions and for skin cancer check-ups, we follow Hair et al.'s recommendations [64] for multigroup analysis in partial least squares structural equation modeling (PLS-SEM). Thus, we applied a bootstrapping procedure and permutation tests with 5000 replicates for multigroup analysis [64]. We analyzed reflective and formative measurement models in a first step before examining our structural models in a second step. First, all reflective variables met the criteria for indicator reliability, as loadings were greater than 0.7. Composite reliability was confirmed, since scores were above 0.7 and average variance extracted was greater than 0.5 [64]. Discriminant validity using the heterotrait-monotrait was also confirmed as scores were less than 0.9 [65]. Second, the formative variables were tested. Convergent validity was present as outer weights and outer loadings were greater than 0.7. Multicollinearity was not given, as the variance inflation factors (VIF) were below 5 [64]. The relative and absolute significance of the indicators was examined by testing and confirming the significance of the weights and loadings. Third, we investigated our structural model concerning effect sizes in behavioral research according to Cohen [66]. Hence, f^2 -values of 0.02, 0.15, and 0.35 should be considered as small, medium, and large effect sizes. Kock and Hadaya [67] state that effect sizes are dependent on sample size and should thus be adjusted according to the sample sizes used. Hence, we determined our effect sizes using G*Power 3.1.9.6 [68]. For both of our models, we determined a f^2 -value of 0.02. To report on explanation powers of our models, we use recommendations of Hair et al. [64], who state that R^2 -values of 0.25, 0.50, and 0.75 represent small, medium, and high explanation powers. To describe the predictive powers of our models, we follow recommendations of Hair et al. [64] using PLS Predict and CVPAT as measures. Whereas PLS Predict considers early antecedents when determining a model's predictive power, CVPAT uses late antecedents instead. Concerning PLS Predict, our prediction errors for both, antibiotics and skin cancer, were not highly symmetrically distributed and we thus report on PLS MAE. We then checked whether values of PLS-SEM MAE were lower

than LM MAE values. Hence, we determined a medium predictive power for skin cancer and no predictive power for antibiotics model. Thus, future predictions of our antibiotics model do not have prediction power concerning early antecedents, i.e., the variables fear (disease) and fear [69]. Concerning CVPAT, we investigated whether the naive indicator-averages as well as prediction benchmarks were surpassed. Hence, we determined that for both models all benchmarks were surpassed and our PLS path models thus have a strong predictive validity. To report on invariances of our model and on significance between groups, we apply the MICOM procedure of Hair et al. [64], whose results are shown in table D.2.

3. Results

3.1. T2DM and skin cancer

The cross-validation comparison revealed a significant performance difference ($p < 0.004$) between the T2DM models trained with and without GNUS augmentation, with the augmented model achieving a higher mean AUC of 0.774 compared to 0.722. The systematic hyperparameter optimization of GNUS identified the optimal scaling factors as $\mu = 0.0001$ and $\sigma = 0.001$. No significant performance improvement was observed when increasing the number of DeepNetBlocks beyond 4 ($p > 0.232$), with four DeepNetBlocks yielding the best mean AUC.

Fig. 3 presents the Receiver Operating Characteristics (ROC) curve for the model with this optimal parameter configuration, achieving an AUC of 0.774 ± 0.030 . When trained on the full dataset, the model achieved its highest AUC on the validation set after 418 epochs. On the test set, the trained model reached an AUC of 0.729. Following training, the model was calibrated using GUESS, which provides physicians with calibrated probability estimates. The resulting framework achieved an accuracy of 0.835 on the test set, with a specificity of 0.90 and a sensitivity of 0.30, based on a decision threshold of 70 % probability for diabetes risk. The corresponding confusion matrix is presented in Appendix B.

The top three performing models in our comparison were InceptionV3, DenseNet169, and DenseNet201. Their performance, measured by the AUC for each skin lesion category, is presented in Table 4. Among them, DenseNet169 achieved the highest overall accuracy of 0.836 and also outperformed the others in terms of weighted F1 score and MCC. A comprehensive comparison of all evaluated models using various

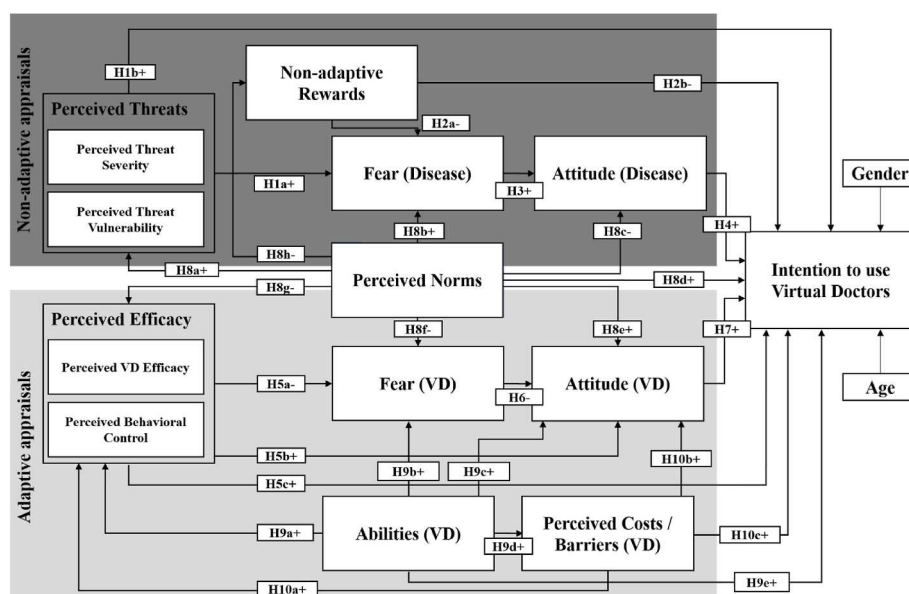


Fig. 2. The integrated theoretical model defined in Ref. [63]. It comprises ten primary nodes representing the driving factors behind the acceptance or rejection of using the Virtual Doctor. The nodes are causally interconnected and the validity of these connections was assessed in the analysis.

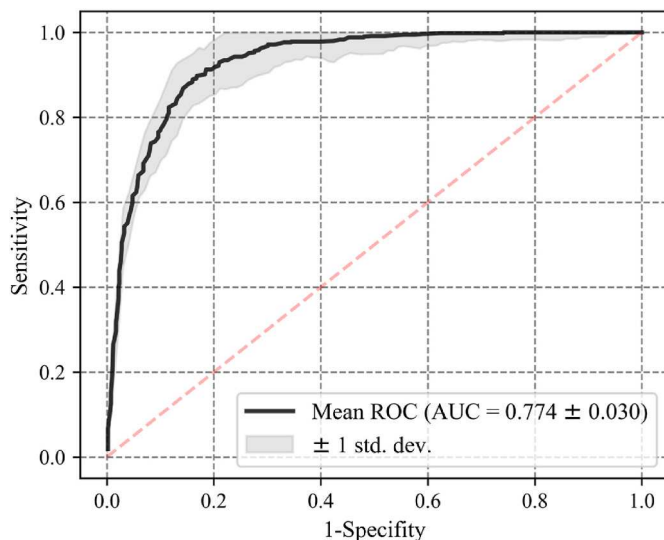


Fig. 3. ROC curve for the dataset using Gaussian noise upsampling followed by stratification. We used the scaling factors $\mu = 0.0001$ and $\sigma = 0.001$.

Table 4

The table shows the best three models for our optimization scheme trained on the HAM10000 data set. The AUC for each class was calculated using the test set for final evaluation.

Architecture	akiec	bcc	bkl	df	mel	nv	vasc
DenseNet169	0.74	0.84	0.85	0.77	0.79	0.88	0.83
DenseNet201	0.82	0.87	0.84	0.78	0.76	0.88	0.80
InceptionV3	0.77	0.84	0.82	0.77	0.77	0.85	0.81

performance metrics is provided in [Appendix B](#).

We integrated DenseNet169 into the Virtual Doctor framework to enhance clinical assistance capabilities. [Fig. 4](#) illustrates an instance of skin lesion evaluation. The recorded image is classified into one of seven skin lesions which cover more than 95 % of all cases in clinical practice [53].

3.2. Integrated model adapted to the Virtual Doctor

Our results reveal that for antibiotics, non-adaptive appraisals were not relevant to determine the intention to use Virtual Doctors. However, the intention to use Virtual Doctors was positively influenced by adaptive appraisals. They reflect that a positive attitude towards Virtual Doctors, which is affected by fear towards Virtual Doctors, perceived norms, abilities to use Virtual Doctors, and perceived efficacy of Virtual

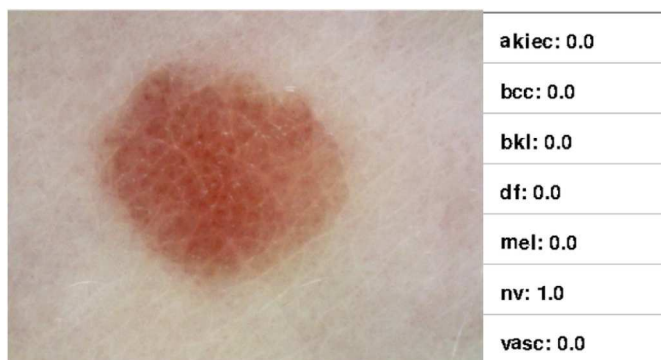


Fig. 4. The figure presents a sample recording for skin cancer check-ups within the Virtual Doctor framework. Classification is determined by the highest value, which corresponds to the output of a softmax function.

Doctors, leads to a higher intention to use them. Further, perceived norms influence the intention to use Virtual Doctors for antibiotics prescriptions and we determined a mean intention value of 4.15 (SD = 1.05) on a 7 point Likert-Scale. The results of the path coefficients and significance of the hypothesis tests are listed in [Appendix D.2](#). For skin cancer check-ups, we also determined that adaptive appraisals are important while non-adaptive appraisals are not. Specifically, the intention to use Virtual Doctors is influenced by attitude towards Virtual Doctors, which in turn is influenced by perceived norms and perceived efficacy as well as abilities to use Virtual Doctors. For skin cancer check-ups, we determined a mean intention value of 4.18 (SD = 1.02). The final interconnections between the significant variables ([Fig. 2](#)) are shown in [Fig. 5](#).

4. Discussion

In our study, we describe and analyze the Virtual Doctor, a semi-autonomous, AI-based decision-support system for automated diagnostics in hospitals, taking experiments and questionnaires for evaluation into account. The system incorporates biomedical sensors configured as a distributed network, ensuring scalability and adaptability to future technological advancements. Currently, these sensors facilitate rapid diagnostics for conditions such as pyrexia, hypoxemia, and mental well-being by directly analyzing sensor data. This capability may be particularly valuable in epidemic scenarios by identifying characteristic indicators and detecting local outbreaks. Furthermore, the system integrates a decision-support module leveraging deep learning algorithms to analyze data related to T2DM and skin cancer, enhancing diagnostic accuracy and clinical decision-making.

We analyzed two use cases as examples, namely the T2DM prediction and skin cancer screening. For the experiments and our T2DM prediction model, we achieved an AUC of 0.729 on the test set, which only incorporated non-invasive data measurable by the Virtual Doctor. While other studies achieved better scores up to an AUC of 0.950, they rely on invasive biomarkers such as insulin level or the HbA1c measurements [70,71]. In the study conducted by Wang et al., an early warning system for T2DM was developed using features that overlap with our approach, such as age, weight, and height [72]. However, their model also incorporated additional variables like family history, which is known to significantly enhance predictive performance. While their system achieved a higher predictive accuracy, with the best AUC reaching 0.913, it relied on disease-specific and targeted questions during data collection.

In contrast, our model is integrated into the Virtual Doctor as a passive warning system. It alerts physicians when the non-invasive biomarkers collected indicate a potential risk of diabetes, without requiring a targeted anamnesis. This passive design makes the model particularly suitable for deployment in the Virtual Doctor framework, enabling unobtrusive and automated risk assessment in unsupervised settings.

Skin cancer classification, our second use case example, requires targeted input from the users. The outcomes of the models for skin lesion classification align with findings reported by Khan et al. [73]. Their study employed more sophisticated feature extraction techniques, achieving an accuracy of 0.804 for DenseNet and 0.840 for InceptionV3, hence slightly outperforming our DenseNet169 by a difference of 0.04. Another study by Yusuf et al. reported a higher accuracy of 89 % [74]. However, the employed EfficientNetV2 architecture demonstrated a lower F1 score of 80 %. Similarly, a study by Gajera et al. compared multiple deep learning models on skin lesion datasets including the ISIC challenge and identified DenseNet as the most effective architecture for the task [75]. Their reported accuracy values are consistent with the findings of our study.

While our results were compared to state-of-the-art networks such as EfficientNetV2 [76], the rapid evolution of machine learning continues to introduce new architectures approaches, including Vision Transformers [77] and multimodal alignment networks [78]. Our approach

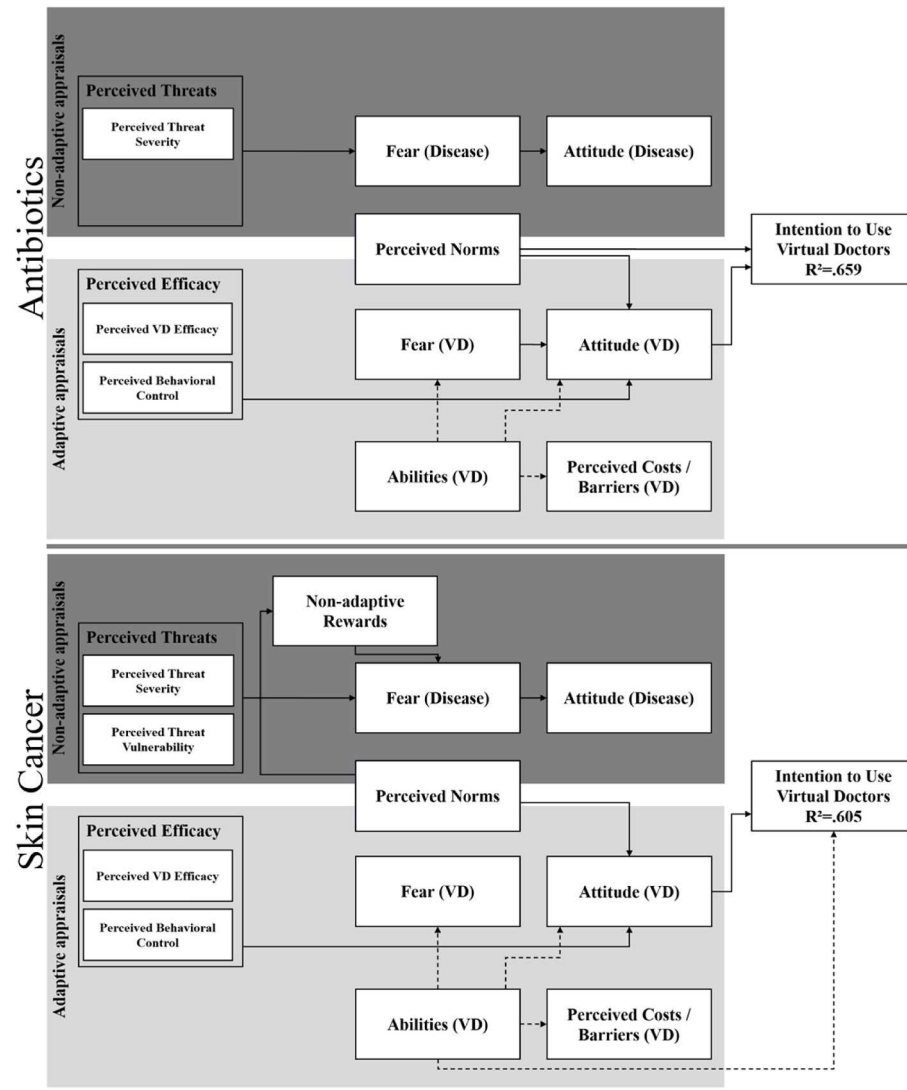


Fig. 5. The figure shows the resulting connections between the driving factors for the patient's intention to use the Virtual Doctor. For the antibiotics prescription case only the attitude towards the Virtual Doctor has an impact on the intention to use. In contrast, the intention to use is directly influenced by the non-adaptive rewards, the attitude towards the Virtual Doctor and the abilities of the Virtual Doctor for the skin-cancer check-up use case.

serves as a baseline that will be refined over time by incorporating future advancements from the scientific community. The implementation aims to enable convenient regular check-ups, empowering users to monitor their health independently. Skin cancer is the most prevalent form of cancer, with cases rising significantly, and early diagnosis significantly improves treatment success rates. By offering skin cancer screenings without long waiting times or financial barriers, this solution has the potential to increase early-stage diagnoses and ultimately save lives [79–81]. Although the integrated neural networks demonstrate potential for supporting clinical decision-making, they have not yet undergone clinical validation. Their current evaluation is based solely on performance within the provided test datasets. In the next phase of the project, the complete framework will be tested in a clinical setting at the hospital in Bochum, and the results will be reported.

The commissioned survey allows us to evaluate the integrated model on the Virtual Doctor use cases. We determined R^2 -values of 0.659 (antibiotics prescription) and 0.605 (skin cancer check-ups) that represent moderate explanatory powers of our models, according to Hair et al. [64]. Compared to related work, our models performed quite similar, as e.g., Baudier et al. [15] determined a R^2 -value of 0.712 in their study investigating the use of telemedicine cabins in France by applying Unified Theory of Acceptance and Use of Technology (UTAUT2) theory.

In contrast, our model not only takes the treatment in question into account, on which UTAUT2 mainly focuses, but also considers the disease in question. Nonetheless, our results support propositions of Baudier et al. [15] in that performance expectancy, called perceived efficacy in our study, is important for the intention to use. From the information retrieved from the integrated we conclude that, first of all there is a general tendency to use the Virtual Doctor. Secondly, it is essential to furnish users with educational materials regarding the disease and the diagnostic assistance provided by Virtual Doctor for check-ups of severe diseases, aiming to alleviate apprehension. We will track the improvements of the Virtual Doctor with the questions 16, 17, and 18 of the designed questionnaire.

The Virtual Doctor is a research prototype in autonomous clinical decision-support systems, with acknowledged limitations in its current developmental stage. While planned for future transition into a medical device, it currently does not comply with the Medical Device Regulations (MDR) in the European Union as Software as a Medical Device (SaMD) nor with the Food and Drug Administration (FDA) regulations for the U.S. market.

Ensuring safety, performance, and reliability in clinical settings requires compliance with regulatory standards, including implementation of standards such as ISO 13485 for Quality Management, ISO 14971 for

Risk Management, IEC 62304 for Software Life Cycle, and IEC 62366 for Usability Engineering. Additionally, a final prospective, clinical evaluation concludes the device's safety and effectiveness, followed by an extensive audit to verify that all required processes were followed according to the standards and regulatory requirements, which is, however, not part of the current research project and is planned in the future.

Future efforts will focus on meeting these standards and collaborating with regulatory experts. Given the importance of ethical considerations, the use of the Virtual Doctor prototype in clinical environments is conducted under strict supervision to ensure patient safety and data security.

Currently, the design and evaluation of the Virtual Doctor are based on a limited number of integrated use cases, which constrains its overall applicability. The ongoing goal of the project is to collect additional data, enabling the expansion of use cases and facilitating further evaluation of social acceptance in real clinical environments with actual patients.

With the increasing automation of medical processes through artificial intelligence, ethical considerations are critical. To ensure maximum data security, the Virtual Doctor operates without internet connectivity, minimizing the risk of external attacks or data breaches. All internal data exchanges are encrypted, and access is restricted through password protection at both the local network level and the database level. In addition, potential bias arising from population selection should be taken into account. The project is also subject to review by an independent ethics committee to ensure adherence to ethical standards regarding data privacy and user protection.

The data generated by the project is intended for use in future research, particularly to enhance the performance of the implemented machine learning architectures. Since the data is stored locally at the hospital without internet connectivity, periodic manual retrieval is necessary. This ensures data protection while also creating opportunities to collaborate with local physicians on potential improvements. Although external access is restricted, physicians can access the data at any time via the local network. Retrieved data may be shared with other researchers upon reasonable research request. A core advantage of the system lies in its cost efficiency. Through custom-designed hardware components, the fully equipped cabin can be produced at an estimated cost of 5700€, making it suitable for deployment across various economic settings. The most expensive components are the cabin construction materials, the medical-grade dermatoscope and the central computing unit. Maintenance is currently handled manually by the research team, with software updates applied onsite when hardware changes are made. As the cabin does not rely on internet connectivity, remote updates are not possible. Ongoing operational costs primarily depend on power consumption, space allocation, and cabin usage. Support for physical infrastructure (e.g., space and power supply) is currently provided by the University Hospital Bochum.

Future developments aim to incorporate more advanced hardware, such as high-performance GPUs, to enable more complex AI functionalities. Planned expansions include integrating an AI-powered chatbot for targeted anamnesis and a fully autonomous fundus camera for ophthalmological assessments.

Regarding ethical and legal considerations, potential liability for misdiagnosis remains a critical concern. Since the system is currently designed as a physician-support tool, all diagnostic decisions remain under medical supervision. However, as the system evolves toward greater autonomy, clear legal frameworks will need to be established to define accountability in case of errors. Furthermore, patient data is stored locally and transmitted only within a secured local network, minimizing risks of unauthorized access or external breaches. These measures, combined with future regulatory compliance efforts, aim to uphold both clinical safety and ethical responsibility.

Despite the relatively low implementation cost, the potential benefits are substantial. The Virtual Doctor is designed to improve the experience of both patients and healthcare professionals. By conducting an initial

health assessment prior to physician-led anamnesis, the system can provide valuable information and decision support, thereby streamlining consultations and improving access to regular check-ups. Ultimately, the Virtual Doctor represents a novel research initiative exploring the deployment of an AI-supported anamnesis framework in real-world settings.

Several commercial and research-based systems exist in the space of digital diagnostics and clinical decision support, such as Babylon Health, Ada Health, and Butterfly iQ. Babylon Health and Ada Health primarily focus on AI-powered symptom checkers, which operate through mobile or web-based interfaces and rely heavily on user-reported information. These tools are typically limited to questionnaire-driven interactions and are not integrated with physical examination hardware or real-time biosignal acquisition. Butterfly iQ, on the other hand, represents a portable ultrasound imaging device that supports point-of-care diagnostics but requires clinical expertise to operate and interpret results.

In contrast, the Virtual Doctor system is designed as an autonomous, cabin-based diagnostic platform that integrates multiple non-invasive sensing modalities (e.g., dermatoscopy and biosignals) into a single physical environment. Unlike most commercial systems, it is fully offline-capable, ensuring high data security and suitability for deployment in low-resource or infrastructure-limited settings. Additionally, it is constructed using custom hardware at low cost making it more accessible than high-end commercial solutions. Finally, the Virtual Doctor emphasizes physical interaction and passive data acquisition, offering a more comprehensive and user-independent diagnostic pathway. These characteristics collectively distinguish our system from existing solutions and underscore its potential for broader global health applications.

5. Conclusion

In this study, we introduced a non-invasive clinical decision-support system, equipped with multiple sensors and designed for the use without additional supervision. It possesses the potential to ease workforce burdens by delivering prompt results alongside AI assistance for diagnostic support. This is particularly crucial in rural areas where a shortage of medical workforce is imminent. The Virtual Doctor can also be viewed as a complement to existing methods like Telemedicine, offering the additional capability of non-invasive biomarker measurements. The Virtual Doctor is a dynamic framework that adapts to emerging technological advancements and enhances its existing capabilities. Through the two use case examples (T2DM and skin cancer classification), we demonstrated that non-invasive biomarkers can aid in the unsupervised detection of severe diseases. Regular check-ups play a crucial role in identifying undetected diseases and preventing their progression to a severe stage. Our study indicated a widespread demand for continuous Virtual Doctor check-ups, such as those for skin cancer. On the other hand, a strong inclination exists towards the intention to use the Virtual Doctor for rapid diagnostic procedures and subsequent prescriptions. For instance, it may be used for diagnosing viral infections, where sensory results such as temperature and medical history serve as the established diagnostic procedures.

In the future, we aim at evaluating the integrated model and refine the human-machine-interface based on user feedback in an evaluation study conducted at the University Hospital Bochum. Moreover, we will enhance the existing framework with additional biomedical sensors and try to increase the number of use cases by analyzing other datasets and exploring new deep learning approaches. The data obtained from the Virtual Doctor will be used for future research endeavors and to enhance the developed models. The outlined data acquisition system will be essential for future researchers to understand and use the Virtual Doctor database.

CRediT authorship contribution statement

Jan Benedikt Ruhland: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Johannes Wichmann:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dmitry Degtyar:** Writing – review & editing, Validation, Software, Methodology, Formal analysis, Data curation. **Roman Martin:** Writing – review & editing, Validation, Resources, Investigation, Conceptualization. **Leon Fehse:** Writing – review & editing, Resources, Investigation. **Jan Henric Klau:** Writing – review & editing, Resources, Investigation. **Thorsten Papenbrock:** Writing – review & editing, Validation, Methodology, Investigation, Funding acquisition. **Jan-Peter Sowa:** Writing – review & editing, Investigation, Conceptualization. **Ali Canbay:** Writing – review & editing, Investigation, Conceptualization. **Bernd Freisleben:** Writing – review & editing, Validation, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Michael Leyer:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dominik Heider:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Ethics statement

All research was conducted in accordance with institutional guidelines and in compliance with German laws.

Data Availability Statement

The results regarding the T2DM prediction model are publicly available on Github <https://github.com/JanRland/Diabetes.git>. The code for generating and training on the HAM10000 dataset for skin lesion classification can be accessed on <https://github.com/JanRland/SkinScreening.git>. The scenarios, questionnaires and data analyses for our commissioned study are available here: https://osf.io/dqjef/?view_only=53682a2c35ec45c88c6cbbbf5a8fc32d.

Appendix A. Virtual Doctor Web Interface and Hardware

The locally hosted website is developed using PHP. It allows control of the ventilation, lighting, and waiting music. New patients must be registered in the database before an autonomous examination can be initiated.

Ethics statement

The authors consciously assure that for the manuscript “Deep Anomaly Detection Enforced Functional Localization Using Fundus Images” the following is fulfilled.

- 1) This material is the authors’ own original work, which has not been previously published elsewhere.
- 2) The paper is not currently being considered for publication elsewhere.
- 3) The paper reflects the authors’ own research and analysis in a truthful and complete manner.
- 4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
- 5) The results are appropriately placed in the context of prior and existing research.
- 6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
- 7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

Funding

The project was partially funded by the Hessian Ministry of Science and Research, Arts and Culture (HMWK) and Hessian.AI.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We express our gratitude to the Hessian Ministry of Science and Research, Arts and Culture (HMWK) and Hessian.AI with its Connectom Program for the financial support of the project Virtual Doctor. Additionally, we extend our thanks to the Marburger Rechen-Cluster 3 (MARC3) for providing the necessary computational resources.

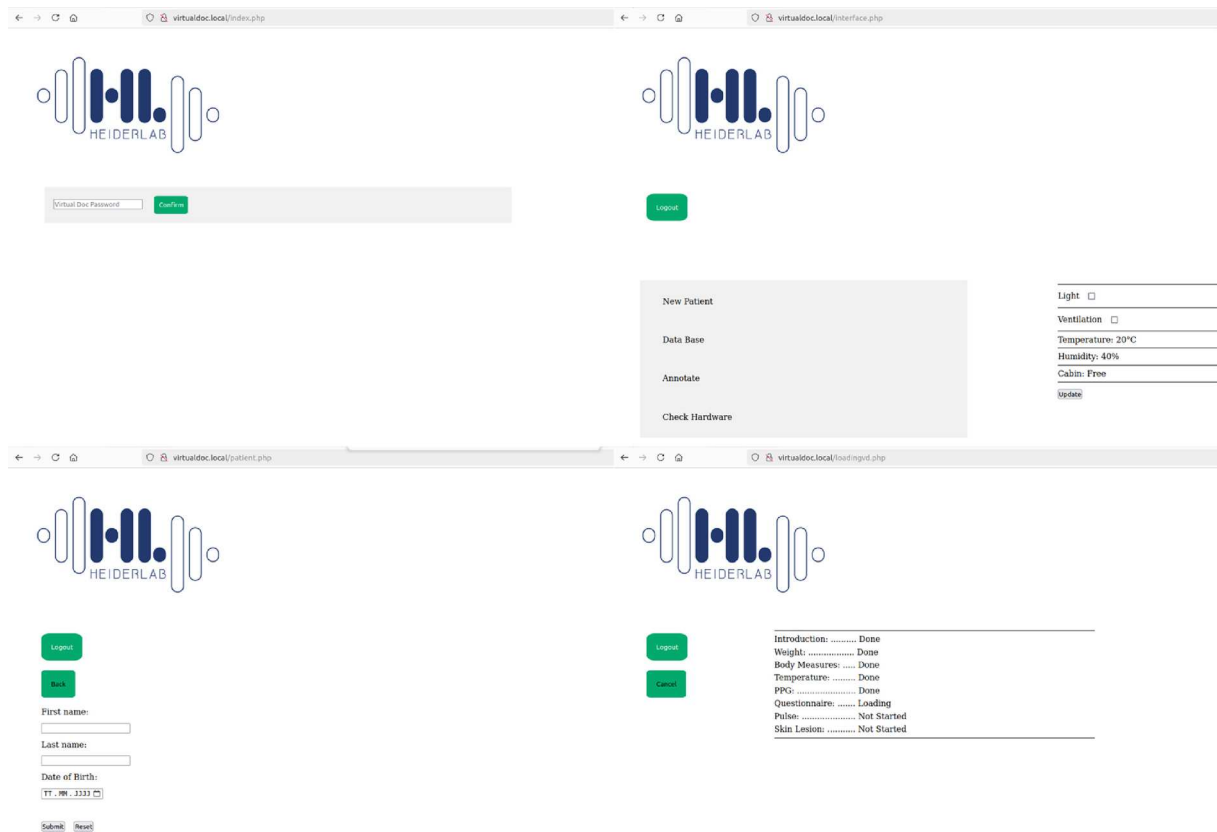


Fig. A.6. The subfigures show images of the local Virtual Doctor website. Access is only possible using the network password and the website password.

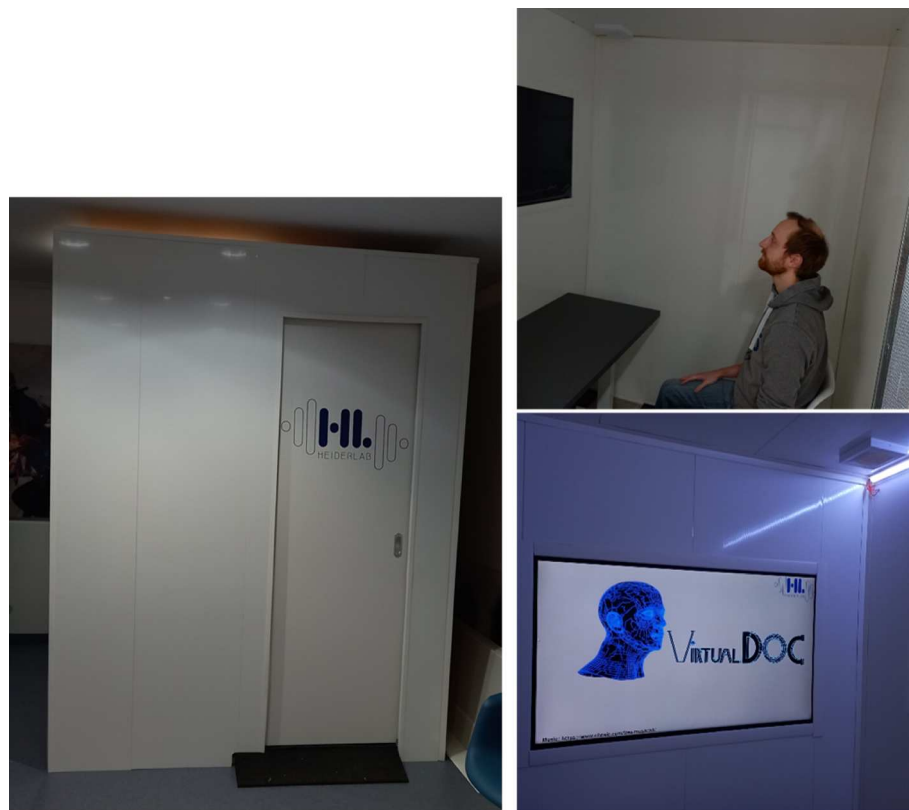


Fig. A.7. The image on the left shows the exterior of the Virtual Doctor located at a German hospital, while the images on the right depict the interior of the Virtual Doctor.

Appendix B. T2DM and Skin Lesion Prediction

Using a 70 % probability threshold for diabetes classification results in a sensitivity of 30 %. The corresponding confusion matrix for the test set is shown in [table B.3](#). The physician will receive GUESS-calibrated probabilities and, as the final diagnosis relies on their judgment, considering additional key indicators such as increased polyuria, the system serves as a passive alert mechanism.

Table B.3

Confusion matrix for the diabetes prediction model on the test set for a decision threshold allowing 10 % missclassification.

	Diabetes (Prediction)	Healthy (Prediction)
Diabetes	16	39
Healthy	40	383

We trained eleven CNN models for the skin lesion classification task. In [table B.4](#), the results for all networks on the validation set are presented, while [table B.5](#) lists the results on the test set.

Table B.4

Results for the machine learning architectures for skin cancer classifications on the validation set.

Architecture	Accuracy	Weighted F1	MCC
DenseNet169	0.906	0.904	0.815
DenseNet201	0.916	0.914	0.837
Inception_V3	0.905	0.902	0.813
ResNet50	0.904	0.900	0.811
ResNet101	0.903	0.899	0.811
ResNet18	0.905	0.902	0.813
ResNet34	0.898	0.894	0.800
DenseNet121	0.916	0.915	0.838
VGG19	0.869	0.869	0.749
SqueezeNet	0.817	0.823	0.673
AlexNet	0.837	0.841	0.699

Table B.5

Results for the machine learning architectures for skin cancer classifications on the test set.

Architecture	Accuracy	Weighted F1	MCC
DenseNet169	0.836	0.831	0.717
DenseNet201	0.835	0.828	0.714
Inception_V3	0.829	0.821	0.701
ResNet50	0.821	0.815	0.691
ResNet101	0.819	0.814	0.690
ResNet18	0.817	0.813	0.686
ResNet34	0.815	0.808	0.679
DenseNet121	0.810	0.804	0.673
VGG19	0.772	0.768	0.610
SqueezeNet	0.750	0.757	0.597
AlexNet	0.723	0.724	0.549

[Table B.6](#) lists the hyperparameter configurations used for training the skin cancer and T2DM prediction models. For the T2DM prediction task, we utilize a custom neural network composed of four DeepNetBlock modules. Each DeepNetBlock includes three linear layers, two adaptive normalization layers [55], and internal skip connections. If necessary, dimensionality adjustment is handled automatically within the block via a learnable linear transformation. Each DeepNetBlock is treated as a single hidden layer in the network. The architecture is preceded by an input normalization step and an initial linear projection with ReLU activation. After passing through the four residual blocks, the network concludes with a fully connected output layer followed by a sigmoid activation to produce a probability score for T2DM risk.

Table B.6

Model Configuration Summary.

Configuration	T2DM	Skin Cancer
Architecture Type	Custom deep residual MLP	Pre-defined CNNs
Layers/Structure	One AdaptiveNorm [55] One linear layer (ReLU) Four DeepNetBlocks (each: $3 \times$ Linear, $2 \times$ AdaptiveNorm, skip connections) One output linear layer (Sigmoid)	Final classification head added, pre-trained weights
Activation Functions	ReLU (hidden), Sigmoid (output)	Softmax (output layer)
Optimizer	SGD	SGD

(continued on next page)

Table B.6 (continued)

Configuration	T2DM	Skin Cancer
Learning Rate (LR) Strategy	LR = 0.03, halved if no loss improvement in 50 epochs	LR = 0.01, annealing over 50 epochs
Loss Function	Binary cross-entropy	Cross-entropy
Epochs	2000	50
Batch Size	None	128
Data Augmentation	GNUS upsampling ($\mu = 0.0001$, $\sigma = 0.001$)	Center crop, rotation, normalization; class-balanced upsampling
Dataset Split	90 % train/10 % test (HNR study)	HAM10000 (train/val), ISIC-2018 (test)
Calibration/Postprocessing	GUESS probabilistic calibration	None

Table B.7 shows the confusion matrix for the DenseNet169 used in the Virtual Doctor framework. The test set was used to create the confusion matrix.

Table B.7

Confusion matrix for the DenseNet169.

	akiec	bcc	bkl	df	mel	nv	vasc
akiec	38	4	1	0	1	5	0
bcc	7	78	5	0	1	8	1
bkl	11	6	134	3	7	42	0
df	4	2	0	21	0	14	0
mel	7	3	9	0	67	71	2
nv	4	5	1	2	9	905	1
vasc	1	1	0	2	0	10	21

Appendix C. Scenarios and Questionnaire

Skin Cancer

Imagine you have discovered a strange looking mole on your upper arm. Knowing that this may be a sign of skin cancer, you want to have this mole examined as soon as possible. You call your dermatologist, who can offer you an appointment in 3 months. As an alternative, there is the Virtual Doctor (VD), from whom you can be examined in 2 h.

Antibiotics

You have had a general feeling of illness, cough, fever and shortness of breath for days. These are signs of severe pneumonia. You urgently need antibiotics against it, which, however, is prescription. You live in a remote area and you have to drive 30 km to your family doctor. There is a Virtual Doctor (VD) in your town, 1 km away.

Appendix C.2. Questionnaire

The questionnaire shown in this appendix contains all questions that remained in our analysis after performing several tests for the measurement and structural models presented in subsection "Evaluation methods". Our full questionnaire that also contains questions that did not pass the tests is shown in the additional materials to this study in subsection "Data Availability Statement".

Intention to use: (1) I would definitely use the described Virtual Doctor if it would be available; (2) I intend to use the described Virtual Doctor if it would be available; (3) I plan to use the described Virtual Doctor if it would be available.

Attitude (VD): The described Virtual Doctor would be ... (1) ... advantageous; (2) ... satisfactory; (3) ... important; (4) ... enjoyable; (5) I would like the use of the described Virtual Doctor.

Fear (VD): The described Virtual Doctor makes me ... (1) ... feel anxious, uncomfortable, as it could hurt me; (2) ... feel anxious, very strained, as I am in danger; (3) ... feel anxious, shaky and nervous.

Perceived VD efficacy: (1) After using the described Virtual Doctor, I think that my sufferings will be reduced; (2) Due to the described Virtual Doctor, my sufferings will be reduced; (3) I am convinced that the described Virtual Doctor will reduce my sufferings.

Perceived behavior control: (1) Whether I use the Virtual Doctor or not is solely dependent on myself; (2) If I really want to, I can use the Virtual Doctor at any time; (3) Whether I use the Virtual Doctor or not, is solely up to me.

Attitude (disease): Having the described disease would be ... (1) ... not advantageous; (2) ... not satisfactory; (4) ... not enjoyable; (5) I would not like having the described disease.

Fear (disease): Having the described disease lets me ... (1) ... feel anxious, uncomfortable, as it could hurt me; (2) ... feel anxious, very strained, as if I am in danger; (3) ... feel anxious, shaky and nervous.

Perceived vulnerability: (1) It is very likely that I fall ill with the disease; (2) My chances to fall ill with the disease are high; (3) The probability that I fall ill with the disease is much higher than within my peer group.

Perceived severity: (2) The disease would cause me major difficulties; (3) When I fall ill with the disease, I get seriously ill; (4) The disease will make me anxious.

Non-adaptive rewards: (2) The waiving of the described Virtual Doctor will spare me pains.

Perceived norms: (1) Individuals from whom I let myself be influenced would advise me to use the described Virtual Doctor; (2) Individuals who are important to me would advise me to use the described Virtual Doctor; (3) Individuals whose opinion I appreciate would advise me to use the described Virtual Doctor; (4) Individuals in a situation comparable to myself would advise me to use the described Virtual Doctor.

Perceived costs/barriers (VD): (1) Using the Virtual Doctor will decrease my costs for diagnosis; (3) Using the Virtual Doctor will decrease my search costs for a fitting doctor; (4) Using the Virtual Doctor will decrease my travel costs.

Abilities (VD): (1) When using the Virtual Doctor, only a short preliminary discussion is necessary with me; (2) I imagine the process of diagnosis by the Virtual Doctor as comfortable; (3) I would have few follow-up questions about the evaluation by the Virtual Doctor.

Appendix D. Hypotheses for the Integrated Model

Appendix D.1. Definition of the Hypotheses

Following the integrated theoretical model, we derive our hypotheses as follows. First, we investigate the health threats and relating appraisals an individual perceives towards the disease [82]. Then, these appraisals influence fear an individual perceives towards the disease [68]. Fear is an emotion leading to anxious or emotional responses when a specific fear-inducing behavior is performed [83]. Hence, we hypothesize that individuals who are at risk of coming down with the disease will (a) feel the emotion of fear more strongly [68] and (b) have a higher intention to use VD for treatment to minimize or avert further negative health outcomes [68,83]. Consequently, our first hypothesis is:

H1. Perceived threats have (a) a positive influence on fear (disease) as well as (b) a positive influence on intention to use. Non-adaptive rewards as benefits individuals receive by refraining the treatment in question may influence fear, as more benefits lower fear towards the disease [30]. Additionally, non-adaptive rewards may negatively influence the intention to use, as receiving more benefits could lead to preventing the use of VD [30]. Thus, our second hypothesis is:

H2. Non-adaptive rewards have a negative influence on (a) fear (disease) and (b) intention to use. Further, attitudes are preferences that are positively or negatively associated with a particular disease [84,85]. Fear as an emotion is known to negatively influence attitudes towards the disease [82,83]. Hence, our third hypothesis is:

H3. Fear (disease) has a positive influence on attitude (disease). Since attitudes towards the disease are a major driver of intention to use [84,85], negative attitudes towards the disease will positively influence intention to use VD. Therefore, our fourth hypothesis is:

H4. Attitude (disease) has a positive influence on intention to use. Second, we consider adaptive appraisals and focus on our treatments, i.e., VD. First, perceived efficacy concerning the intervention investigates whether the treatment is useful to address a specific disease [82,86,87]. For healthcare, perceived VD efficacy is one of the most important elements to determine the intention to use new technologies [87,88]. Further, perceived efficacy is influenced by perceived behavioral control which may also influence the intention to use [84,88]. Then, perceived efficacy influences fear towards the treatment, i.e. VD, as PMT states [82], as well as attitude towards the VD, as UTAUT [87] and RAA [84] state. Hence, our fifth hypothesis is:

H5. Perceived efficacy has a negative influence on (a) fear (VD) as well as a positive influence on (b) attitude (VD) and (c) intention to use. In studies addressing technology acceptance, fear (VD) is a key predictor for the attitude towards the VD and ultimately the intention to use it, e.g., by considering privacy concerns [15,87]. Thus, our sixth hypothesis is:

H6. Fear (VD) has a negative influence on attitude (VD). Accordingly, attitude (VD) is important to investigate the intention to use VD [84,89], as attitude towards new technologies is the major predictor for intention to use [84]. Hence, our seventh hypothesis is:

H7. Attitude (VD) has a positive influence on intention to use. In addition to the duality approach, our integrated model investigates whether perceived social norms (e.g., pressure/recommendations to use new technologies) influence the intention to use. In the past, perceived norms have proven to be important to investigate the intention to use HIS [15,87,88]. Thus, perceived norms influence all predictors in the integrated models, while we assume that adaptive appraisals will be positively and non-adaptive appraisals negatively influenced. Hence, our eighth hypothesis is:

H8. Perceived norms have, on the one side, a positive influence on (a) perceived threats, (b) fear (disease), (c) intention to use, (d) attitude (VD), and (e) perceived efficacy, and on the other side, a negative influence on (f) attitude (disease), (g) fear (VD), and (h) non-adaptive rewards.

Further according to RAA, abilities are important when determining the intention to use VD. As such, high abilities in our study reflect that individuals are able to quickly understand the meaning, risks, and specifications of VD treatments, which then increases their intention to use [15,18]. Thus, our ninth hypothesis is:

H9. Abilities towards VD positively influence (a) perceived efficacy, (b) fear (VD), (c) attitude (VD), (d) perceived costs/barriers, and (e) intention. Additionally, according to UTAUT2, costs/barriers that individuals perceive towards the VD are important in their determination of intention to use. If individuals perceive high costs/barriers towards the VD, their intention to use is lower [15]. Hence, our tenth hypothesis is:

H10. Cost/barriers towards VD negatively influence (a) perceived efficacy, (b) attitude (VD), as well as (c) intention.

Appendix D.2. Results for the hypothesis testing

We evaluated our hypotheses by utilizing our integrated theoretical model to determine our results. In table D.5 all path coefficients determined with PLS-SEM are listed representing the significance of the causal link in Fig. 2.

Table D.5

The table presents the path coefficients for both the original model and the specific use case antibiotics prescriptions. We utilized one-tailed tests to ascertain significance.

Hypothesis	Original Difference (Sig.)	Antibiotics (Sig.; T-value; Confidence Interval; f^2 -value)	Skin cancer (Sig.; T-value; Confidence Interval; f^2 -value)
H1a	-0.039 (p = 0.225)	0.544 (p ≤ 0.001; T = 15.565; [0.491,0.606]; f^2 = 0.405)	0.584 (p ≤ 0.001; T = 16.569; [0.528,0.645]; f^2 = 0.514)
H1b	-0.089 (p = 0.009)	-0.061 (p = 0.008; T = 2.391; [-0.102,-0.018]; f^2 = 0.010)	0.028 (p = 0.168; T = 0.962; [-0.019,0.077]; f^2 = 0.002)

(continued on next page)

Table D.5 (continued)

Hypothesis	Original Difference (Sig.)	Antibiotics (Sig.; T-value; Confidence Interval; f^2 -value)	Skin cancer (Sig.; T-value; Confidence Interval; f^2 -value)
H2a	0.164 (p = 0.000)	0.019 (p = 0.291; T = 0.551; [-0.038,0.074]; f^2 = 0.000)	-0.145 (p ≤ 0.001; T = 4.011; [-0.204,-0.084]; f^2 = 0.032)
H2b	0.001 (p = 0.493)	-0.028 (p = 0.127; T = 1.143; [-0.070,0.012]; f^2 = 0.002)	-0.029 (p = 0.154; T = 1.018; [-0.074,0.018]; f^2 = 0.002)
H3	-0.019 (p = 0.363)	0.516 (p ≤ 0.001; T = 13.787; [0.455,0.577]; f^2 = 0.360)	0.534 (p ≤ 0.001; T = 13.594; [0.471,0.599]; f^2 = 0.399)
H4	-0.002 (p = 0.481)	0.023 (p = 0.149; T = 1.041; [-0.014,0.058]; f^2 = 0.001)	0.025 (p = 0.169; T = 0.958; [-0.018,0.066]; f^2 = 0.001)
H5a	-0.031 (p = 0.347)	-0.116 (p = 0.017; T = 2.118; [-0.207,-0.027]; f^2 = 0.009)	-0.085 (p = 0.051; T = 1.635; [-0.173,-0.002]; f^2 = 0.006)
H5b	0.036 (p = 0.248)	0.362 (p ≤ 0.001; T = 9.293; [0.298,0.427]; f^2 = 0.157)	0.326 (p ≤ 0.001; T = 9.412; [0.270,0.384]; f^2 = 0.161)
H5c	0.017 (p = 0.387)	0.090 (p = 0.021; T = 2.034; [0.022,0.165]; f^2 = 0.010)	0.073 (p = 0.039; T = 1.758; [0.005,0.141]; f^2 = 0.007)
H6	-0.068 (p = 0.058)	-0.111 (p ≤ 0.001; T = 3.890; [-0.158,-0.064]; f^2 = 0.024)	-0.043 (p = 0.090; T = 1.339; [-0.097,0.008]; f^2 = 0.004)
H7	0.005 (p = 0.463)	0.605 (p ≤ 0.001; T = 14.405; [0.534,0.672]; f^2 = 0.452)	0.600 (p ≤ 0.001; T = 12.487; [0.517,0.675]; f^2 = 0.369)
H8a	0.001 (p = 0.397)	-0.001 (p = 0.336; T = 0.343; [-0.004,0.003]; f^2 = 0.001)	-0.001 (p = 0.329; T = 0.443; [-0.007,0.004]; f^2 = 0.001)
H8b	0.009 (p = 0.422)	-0.014 (p = 0.326; T = 0.451; [-0.066,0.038]; f^2 = 0.000)	-0.024 (p = 0.244; T = 0.694; [-0.078,0.034]; f^2 = 0.001)
H8c	-0.019 (p = 0.356)	-0.070 (p = 0.020; T = 2.051; [-0.126,-0.014]; f^2 = 0.007)	-0.051 (p = 0.080; T = 1.406; [-0.110,0.009]; f^2 = 0.004)
H8d	0.121 (p = 0.026)	0.182 (p ≤ 0.001; T = 4.045; [0.108,0.255]; f^2 = 0.054)	0.060 (p = 0.068; T = 1.494; [-0.005,0.128]; f^2 = 0.005)
H8e	-0.040 (p = 0.277)	0.286 (p ≤ 0.001; T = 7.680; [0.222,0.345]; f^2 = 0.125)	0.326 (p ≤ 0.001; T = 8.936; [0.266,0.387]; f^2 = 0.179)
H8f	0.030 (p = 0.330)	-0.042 (p = 0.224; T = 0.759; [-0.133,0.049]; f^2 = 0.001)	-0.072 (p = 0.052; T = 1.625; [-0.147,-0.000]; f^2 = 0.004)
H8g	0.001 (p = 0.127)	0.002 (p = 0.037; T = 1.786; [0.001,0.004]; f^2 = 0.008)	0.001 (p = 0.154; T = 1.018; [-0.000,0.002]; f^2 = 0.005)
H8h	-0.047 (p = 0.207)	0.103 (p = 0.008; T = 2.430; [0.031,0.174]; f^2 = 0.011)	0.150 (p ≤ 0.001; T = 3.634; [0.080,0.215]; f^2 = 0.023)
H9a	-0.001 (p = 0.100)	-0.000 (p = 0.336; T = 0.424; [-0.002,0.001]; f^2 = 0.000)	0.001 (p = 0.211; T = 0.804; [-0.000,0.002]; f^2 = 0.004)
H9b	0.019 (p = 0.397)	-0.313 (p ≤ 0.001; T = 6.007; [-0.400,-0.230]; f^2 = 0.072)	-0.332 (p ≤ 0.001; T = 6.897; [-0.411,-0.253]; f^2 = 0.092)
H9c	-0.077 (p = 0.098)	0.173 (p ≤ 0.001; T = 3.813; [0.098,0.248]; f^2 = 0.036)	0.249 (p ≤ 0.001; T = 5.970; [0.180,0.316]; f^2 = 0.087)
H9d	0.090 (p = 0.043)	0.564 (p ≤ 0.001; T = 16.351; [0.506,0.620]; f^2 = 0.467)	0.474 (p ≤ 0.001; T = 12.719; [0.414,0.535]; f^2 = 0.289)
H9e	-0.141 (p = 0.007)	0.007 (p = 0.420; T = 0.202; [-0.051,0.064]; f^2 = 0.000)	0.148 (p ≤ 0.001; T = 3.427; [0.078,0.220]; f^2 = 0.031)
H10a	0.001 (p = 0.223)	0.002 (p = 0.075; T = 1.441; [0.000,0.004]; f^2 = 0.009)	0.001 (p = 0.074; T = 1.443; [-0.000,0.002]; f^2 = 0.010)
H10b	-0.006 (p = 0.464)	0.030 (p = 0.237; T = 0.716; [-0.039,0.100]; f^2 = 0.001)	0.036 (p = 0.153; T = 1.024; [-0.021,0.094]; f^2 = 0.002)
H10c	0.080 (p = 0.057)	0.037 (p = 0.137; T = 1.095; [-0.019,0.091]; f^2 = 0.002)	-0.043 (p = 0.124; T = 1.155; [-0.104,0.020]; f^2 = 0.003)

References

- [1] W. H. O. Regional Office for Europe, Health and Care Workforce in Europe: Time to act, World Health Organization, Regional Office for Europe, 2022.
- [2] W.H.O. Bucharest, Declaration on the health and care workforce. <https://iris.who.int/bitstream/handle/10665/366519/Bucharest-decl-2023-eng.pdf?sequence=1>, 2023, 09.
- [3] J. Campbell, G. Dussault, J. Buchan, F. Pozo-Martin, M. Guerra Arias, C. Leone, A. Siyam, G. Cometto, A Universal Truth: No Health Without a Workforce. Forum Report, Third Global Forum on Human Resources for Health, Global Health Workforce Alliance and World Health Organization, Recife, Brazil, Geneva, 2013.
- [4] C.E. Miller, R.S. Vasan, The southern rural health and mortality penalty: a review of regional health inequities in the United States, Soc. Sci. Med. 268 (2021 Jan) 113443, <https://doi.org/10.1016/j.socscimed.2020.113443>. Epub 2020 Oct 23. PMID: 33137680; PMCID: PMC7755690.
- [5] D. Terry, B. Peck, E. Baker, D. Schmitz, The rural nursing workforce hierarchy of needs: decision-making concerning future rural healthcare employment, Healthcare 9 (9) (2021) 1232, <https://doi.org/10.3390/healthcare9091232>.
- [6] R. Subedi, T.L. Greenberg, S. Roshanfarshar, Does geography matter in mortality? An analysis of potentially avoidable mortality by remoteness index in Canada, Health Rep. 30 (5) (2019 May 15) 3–15, <https://doi.org/10.25318/82-003-x201900500001-eng>. PMID: 31091331.
- [7] W. Yan, G. Sun, Income, workload, and any other factors associated with anticipated retention of rural doctors? Prim. Health Care Res. Dev. 23 (2022) e12, <https://doi.org/10.1017/S1463423621000839>.
- [8] D. Russell, S. Mathew, M. Fitts, Z. Liddle, L. Murakami-Gold, N. Campbell, M. Ramjan, Y. Zhao, S. Hines, J.S. Humphreys, J. Wakerman, Interventions for health workforce retention in rural and remote areas: a systematic review, Hum. Resour. Health 19 (1) (2021 Aug 26) 103, <https://doi.org/10.1186/s12960-021-00643-7>. PMID: 34446042; PMCID: PMC8393462.
- [9] P. Rajpurkar, E. Chen, O. Banerjee, E.J. Topol, AI in health and medicine, Nat. Med. 28 (1) (2022 Jan) 31–38, <https://doi.org/10.1038/s41591-021-01614-0>. Epub 2022 Jan 20. PMID: 35058619.
- [10] A.S. Ahuja, The impact of artificial intelligence in medicine on the future role of the physician, PeerJ 7 (2019 Oct 4) e7702, <https://doi.org/10.7717/peerj.7702>. PMID: 31592346; PMCID: PMC6779111.
- [11] S.T. Rikta, K.M.M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, S.K. Dey, XML-GBM lung: an explainable machine learning-based application for the diagnosis of lung cancer, J. Pathol. Inf. 14 (2023 Mar 24) 100307, <https://doi.org/10.1016/j.jpi.2023.100307>. PMID: 37025326; PMCID: PMC10070138.
- [12] S. Padash, J.P. Mickley, D.V. Vera Garcia, F. Nuge, B.J. Erickson, C. C. Wyles, M.J. Taunton, An overview of machine learning in orthopedic surgery: an educational paper, J. Arthroplast. 38 (10) (2023 Oct) 1938–1942, <https://doi.org/10.1016/j.arth.2023.08.043>. Epub 2023 Aug 19. PMID: 37598786; PMCID: PMC10601337.
- [13] A. Motwani, P.K. Shukla, M. Pawar, Ubiquitous and smart healthcare monitoring frameworks based on machine learning: a comprehensive review, Artif. Intell. Med. 134 (2022 Dec) 102431, <https://doi.org/10.1016/j.artmed.2022.102431>. Epub 2022 Oct 22. PMID: 36462891; PMCID: PMC9595483.
- [14] F. Mair, G. Peterkin, A. Laing, J. Ferguson, S. Fraser, Feasibility of a telemedicine 'booth' for supporting remote care in Scotland, J. Telemed. Telecare 14 (3) (2008) 127–129, <https://doi.org/10.1258/jtt.2008.003007>. PMID: 18430277.
- [15] P. Baudier, G. Kondrateva, C. Ammi, The future of telemedicine cabin? The case of the French students' acceptability, Futures 122 (2020 Sep) 102595, <https://doi.org/10.1016/j.futures.2020.102595>. Epub 2020 Jun 13. PMID: 32834075; PMCID: PMC7293447.
- [16] G. Falgarone, G. Bousquet, A. Wilmet, A. Brizio, V. Faure, C. Guillouet, F. Baudino, I. Roque, S. Mayol, F. Pamoukdjian, A teleconsultation device, consult station, for remote primary care: multisite prospective cohort study, J. Med. Internet Res. 24 (5) (2022 May 17) e33507, <https://doi.org/10.2196/33507>. Erratum in: J Med Internet Res. 2022 Oct 13;24(10):e43220. PMID: 35254263; PMCID: PMC9157322. doi: 10.2196/43220.
- [17] T. Davenport, The Future of Work Now: Good Doctor Technology for Intelligent Telemedicine in Southeast Asia, 2021. (Accessed 18 June 2023).
- [18] A. Haleem, M. Javaid, R.P. Singh, R. Suman, Telemedicine for healthcare: capabilities, features, barriers, and applications, Sens. Int. 2 (2021) 100117, <https://doi.org/10.1016/j.sintl.2021.100117>. Epub 2021 Jul 24. PMID: 34806053; PMCID: PMC8590973.
- [19] Tanja Sophie Gask, Johannes Wichmann, Michael Leyer, Health information systems: potential users balancing adaptive and maladaptive appraisals, in: ICIS 2021 Proceedings, 2021.
- [20] R. Ftouni, B. AlJardali, M. Hamdanieh, L. Ftouni, N. Salem, Challenges of telemedicine during the COVID-19 pandemic: a systematic review, BMC Med. Inf. Decis. Making 22 (1) (2022 Aug 3) 207, <https://doi.org/10.1186/s12911-022-01952-0>. PMID: 35922817; PMCID: PMC9351100.
- [21] L. Hwei, G. Octavius, Potential advantages and disadvantages of telemedicine: a literature review from the perspectives of patients, medical personnel, and hospitals, J. Commun. Empowerment for Health 4 (2021) 180–186, <https://doi.org/10.22146/jcoemph.64247>.
- [22] R. Wang, J. Zhang, S. He, H. Guo, T. Li, Q. Zhong, J. Ma, J. Xu, K. He, Design and application of a novel telemedicine system jointly driven by multinet network integration and remote control: practical experience from PLAGH, China, Healthc Technol. Lett 10 (6) (2023 Dec 5) 113–121, <https://doi.org/10.1049/htl2.12057>. PMID: 38111799; PMCID: PMC10725722.
- [23] L. Wagner, A. Vehorn, A.S. Weitlauf, A.M. Lavanderos, J. Wade, L. Corona, Z. Warren, Development of a novel telemedicine tool to reduce disparities related to the identification of preschool children with autism, J. Autism Dev. Disord. 55 (1) (2025 Jan) 30–42, <https://doi.org/10.1007/s10803-023-06176-3>. Epub 2023 Dec 8. PMID: 38064003; PMCID: PMC11161552.
- [24] Ting Tan, Arun Thirunavukarasu, Liyuan Jin, Joshua Lim, Stanley Poh, Ling Teo, Marcus Ang, Paul Chan, Jasmine Ong, Angus Turner, Jonas Karlström, Tien Yin Wong, Jude Stern, Daniel Ting, Artificial intelligence and digital health in global eye health: opportunities and challenges, Lancet Global Health 11 (2023) E1432–E1443, [https://doi.org/10.1016/S2214-109X\(23\)00323-6](https://doi.org/10.1016/S2214-109X(23)00323-6).
- [25] M. Ullah, S. Hamayun, A. Wahab, S.U. Khan, M.U. Rehman, Z.U. Haq, K. U. Rehman, A. Ullah, A. Mehreen, U.A. Awan, M. Qayum, M. Naem, Smart technologies used as smart tools in the management of cardiovascular disease and their future perspective, Curr. Probl. Cardiol. 48 (11) (2023 Nov) 101922, <https://doi.org/10.1016/j.cpcardiol.2023.101922>. Epub 2023 Jul 10. PMID: 37437703.

- [26] Solovchuk D. Rodoplu, Advances in AI-assisted biochip technology for biomedicine, *Biomed. Pharmacother.* 177 (2024 Aug) 116997, <https://doi.org/10.1016/j.biopha.2024.116997>. Epub 2024 Jun 28. PMID: 38943990.
- [27] Yang Tan, Zhixing Zhang, Mingchen Li, Fei Pan, Hao Duan, Zijie Huang, Hua Deng, Zhuohang Yu, Chen Yang, Guoyang Shen, Peng Qi, Chengyuan Yue, Yuxian Liu, Liang Hong, Huiqun Yu, Guisheng Fan, Yun Tang, MedChatZH: a tuning LLM for traditional Chinese medicine consultations, *Comput. Biol. Med.* 172 (2024) 108290, <https://doi.org/10.1016/j.compbiomed.2024.108290>. ISSN 0010-4825.
- [28] K. Singhal, T. Tu, J. Gottweis, et al., Toward expert-level medical question answering with large language models, *Nat. Med.* 31 (2025) 943–950, <https://doi.org/10.1038/s41591-024-03423-7>.
- [29] T. Tu, M. Schaekermann, A. Palepu, et al., Towards conversational diagnostic artificial intelligence, *Nature* (2025), <https://doi.org/10.1038/s41586-025-08866-7>.
- [30] S. Spänig, A. Emberger-Klein, J.P. Sowa, A. Canbay, K. Menrad, D. Heider, The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes, *Artif. Intell. Med.* 100 (2019 Sep) 101706, <https://doi.org/10.1016/j.artmed.2019.101706>. Epub 2019 Aug 21. PMID: 31607340.
- [31] B. Rozemberczki, The Shapley Value in Machine Learning", arXiv:2202.05594, 2022, <https://doi.org/10.48550/arXiv.2202.05594>.
- [32] M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks", arXiv:1703.01365, 2017, <https://doi.org/10.48550/arXiv.1703.01365>.
- [33] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", arXiv: 1610.02391, 2016, <https://doi.org/10.48550/arXiv.1610.02391>.
- [34] S.E. Aliouane, H. Chehili, K. Boulahrouf, A. Abdelaziz, N. Khelifa, M.A. Hamidechi, Integrating deep learning and SHAP for breast cancer classification and biomarker discovery using gene expression data, *IEEE Access* 13 (2025) 49693–49709, <https://doi.org/10.1109/ACCESS.2025.3552280>.
- [35] P. Narayankar, V.P. Baligar, Explainability of brain tumor classification based on region, in: 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), 2024, pp. 1–6, <https://doi.org/10.1109/ICETCS61022.2024.10544289>. Bengaluru, India.
- [36] S. Falahati, M. Alizadeh, F. Ghazipour, Z. Safahi, N. Khaledian, M.R. Salmanpour, An AI-powered public health automated kiosk system for personalized care: an experimental pilot study, arXiv:2504.13880 (2025), <https://doi.org/10.48550/arXiv.2504.13880>.
- [37] M.J. Iqbal, Z. Javed, H. Sadia, et al., Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future, *Cancer Cell Int.* 21 (2021) 270, <https://doi.org/10.1186/s12935-021-01981-1>.
- [38] J. Grosman, Fine-tuned XLSR-53 large model for speech recognition in German. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>, 2021.
- [39] A. Baeviski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: a Framework for Self-Supervised Learning of Speech Representations, arXiv e-prints, June 2020 arXiv: 2006.11477.
- [40] D.C. Brown, Decentering distortion of lenses. *Photogramm. Eng.* 32 (1966) 444–462.
- [41] J. Weng, P.R. Cohen, M. Herniou, Camera calibration with distortion models and accuracy evaluation, *IEEE Trans. Pattern Anal. Mach. Intell.* 14 (1992) 965–980.
- [42] Thorsten Thormählen, Hellward Broszio, Ingolf Wassermann, Robust line-based calibration of lens distortion from a single view, *Mirage* (2003).
- [43] Z. Zhang, A flexible new technique for camera calibration. *Pattern analysis and machine intelligence*, IEEE Trans. on 22 (2000) 1330–1334, <https://doi.org/10.1109/34.888718>.
- [44] Yan Pan, Daniel Gareau, Alon Scope, Milind Rajadhyaksha, Nizar Mullani, Ashfaq Marghoob, Polarized and nonpolarized dermoscopy: the explanation for the observed differences, *Arch. Dermatol.* 144 (2008) 828–829, <https://doi.org/10.1001/archderm.144.6.828>.
- [45] Fahri Auffer, Muhammad Murti, Muhammad Barri, Design of Non-contact Thermometer Using Thermal Camera for Detecting People with Fever, 2021, pp. 1–5, <https://doi.org/10.1109/IC2SE52832.2021.9792101>.
- [46] Jia-Wei Lin, Ming-Hung Lu, Yuan-Hsiang Lin, A Thermal Camera Based Continuous Body Temperature Measurement System, 2019, pp. 1681–1687, <https://doi.org/10.1109/ICCVW.2019.00208>.
- [47] M. Bergner, R.M. Kaplan, J.E. Ware, Measuring overall health: an evaluation of three important approaches, *J. Chron. Dis.* (1987), [https://doi.org/10.1016/S0021-9681\(87\)80027-9](https://doi.org/10.1016/S0021-9681(87)80027-9).
- [48] Allyson Ross Davies, E. John, Ware. *Measuring Health Perceptions in the Health Insurance Experiment*, RAND Corporation, Santa Monica, CA, 1981.
- [49] C.L.M. Keyes, Social well-being, *Soc. Psychol. Q.* 61 (2) (1998) 121–140, <https://doi.org/10.2307/2787065>.
- [50] C.D. Ryff, Happiness is everything, or is it? Explorations on the meaning of psychological well-being, *J. Pers. Soc. Psychol.* 57 (1989) 1069–1081, <https://doi.org/10.1037/0022-3514.57.6.1069>.
- [51] A. Schmermund, S. Möhlenkamp, A. Stang, D. Grönemeyer, R. Seibel, H. Hirche, K. Mann, W. Siffert, K. Lauterbach, J. Siegrist, K.H. Jöckel, R. Erbel, Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the heinz nixdorf RECALL study. Risk factors, evaluation of coronary calcium and lifestyle, *Am. Heart J.* 144 (2) (2002 Aug) 212–218, <https://doi.org/10.1067/mhj.2002.123579>. PMID: 12177636.
- [52] L. Ismail, H. Materwala, J. Al Kaabi, Association of risk factors with type 2 diabetes: a systematic review, *Comput. Struct. Biotechnol. J.* 19 (2021 Mar 10) 1759–1785, <https://doi.org/10.1016/j.csbj.2021.03.003>. PMID: 33897980; PMCID: PMC8050730.
- [53] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions, *Sci. Data* 5 (2018) 180161, <https://doi.org/10.1038/sdata.2018.161>.
- [54] N. Codella, V. Rotemberg, P. Tschandl, M. Emre Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC), arXiv e-prints, p. arXiv:1902.03368 (Feb. 2019).
- [55] Jan Benedikt Ruhland, Iraj Masoudian, Dominik Heider, Enhancing deep neural network training through learnable adaptive normalization, *Knowl. Base Syst.* 326 (2025) 113968, <https://doi.org/10.1016/j.knsys.2025.113968>. ISSN 0950-7051.
- [56] J. Beinecke, D. Heider, Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making, *BioData Min.* 14 (2021) 49, <https://doi.org/10.1186/s13040-021-00283-6>.
- [57] J. Schwarz, D. Heider, GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making, *Bioinformatics* 35 (14) (2019 Jul 15) 2458–2465, <https://doi.org/10.1093/bioinformatics/bty984>. PMID: 30496351.
- [58] N.K. Pati, Y.S. Asish, K. Manoj Kumar, et al., Oversampled two-dimensional deep learning model for septenary classification of skin lesion disease, *Natl. Acad. Sci. Lett.* 46 (2023) 159–164, <https://doi.org/10.1007/s40009-022-01175-x>.
- [59] K. Nakamura, B. Derbel, K.-J. Won, B.-W. Hong, Learning-rate annealing methods for deep neural networks, *Electronics* 10 (16) (2021) 2029, <https://doi.org/10.3390/electronics10162029>.
- [60] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, J.P. Cunningham, Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep Learning, 2020, <https://doi.org/10.48550/arXiv.2011.05231> arXiv e-prints, Art. no. arXiv: 2011.05231.
- [61] Joseph Goodman, Cynthia Cryder, Amar Cheema, Data collection in a flat world: the strengths and weaknesses of mechanical turk samples, *J. Behav. Decis. Making* 26 (2013) 213–224, <https://doi.org/10.1002/bdm.1753>.
- [62] Ick Aizen, The theory of planned behavior, *Organ. Behav. Hum. Decis. Process.* 50 (1991) 179–211, [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T).
- [63] T.S. Gesk, J. Wichmann, M. Leyer, The disease or the vaccination: which is more important when deciding whether to vaccinate? *Soc. Sci. Med.* 330 (2023 Aug) 116042 <https://doi.org/10.1016/j.socscimed.2023.116042>. Epub 2023 Jun 24. PMID: 37399657.
- [64] Joseph Hair, Jeff Risher, Marko Sarstedt, Christian Ringle, When to use and how to report the results of PLS-SEM, *Eur. Bus. Rev.* 31 (2018) 2018, <https://doi.org/10.1108/EBR-11-2018-0203>.
- [65] Jörg Henseler, Theo Dijkstra, Marko Sarstedt, Christian Ringle, Adamantios Diamantopoulos, Detmar Straub, David Jr, Joseph Hair, G. Tomas M. Hult, Roger Calantone, Common beliefs and reality about PLS: comments on rönkkö and evermann (2013), *Organ. Res. Methods* 17 (2014) 182–209, <https://doi.org/10.1177/1094428114526928>.
- [66] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, second ed., Lawrence Erlbaum Associates, 1988 <https://doi.org/10.4324/9780203771587>.
- [67] Ned Kock, Pierre Hadaya, Minimum sample size estimation in PLS-SEM: the inverse square root and gamma-exponential methods: sample size in PLS-based SEM, *Inf. Syst. J.* 28 (1) (2016) 227–261, <https://doi.org/10.1111/isj.12131>, 2018.
- [68] S. Mayr, E. Erdfelder, A. Buchner, F. Faul, A short tutorial of gpower, *Tutorials Quantitative Methods Psychol.* 3 (2) (2007) 51–59.
- [69] Nicholas Danks, The piggy in the middle: the role of mediators in PLS-SEM-based prediction, *Data Base for Adv. Information Sys.* (2021), <https://doi.org/10.1145/3505639.3505644>.
- [70] J. Kälsch, L.P. Bechmann, D. Heider, J. Best, P. Manka, H. Kälsch, J.P. Sowa, S. Moebus, U. Slomiany, K.H. Jöckel, R. Erbel, G. Gerken, A. Canbay, Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort, *Sci. Rep.* 5 (2015 Aug 13) 13058, <https://doi.org/10.1038/srep13058>. PMID: 26269425; PMCID: PMC4535035.
- [71] Md Kamrul Hasan, Md Ashraf Alam, Dola Das, Eklas Hossain, Mahmudul Hasan, Diabetes prediction using ensemble of different machine learning classifiers, *IEEE Access* (2020), <https://doi.org/10.1109/ACCESS.2020.2989857>, 1–1.
- [72] X. Wang, J. Ren, H. Ren, W. Song, Y. Qiao, Y. Zhao, L. Linghu, Y. Cui, Z. Zhao, L. Chen, L. Qiu, Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta, *Sci. Rep.* 13 (1) (2023 Aug 5) 12718, <https://doi.org/10.1038/s41598-023-40036-5>. PMID: 37543637; PMCID: PMC10404250.
- [73] M.A. Khan, M. Sharif, T. Akram, R. Damaševičius, R. Maskeliūnas, Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization, *Diagnostics* 11 (5) (2021 Apr 29) 811, <https://doi.org/10.3390/diagnostics11050811>. PMID: 33947117; PMCID: PMC8145295.
- [74] M. Franklin Yusuf, I. Wijayanto, F. Akhyar, T. Dewi Rahmawati, R. Satria Firdhaust, EfficientNetV2-Based deep learning classifier for skin cancer lesion on imbalanced dataset, in: 2024 International Conference on Data Science and Its Applications (ICoDSA), 2024, pp. 498–502, <https://doi.org/10.1109/ICoDSA62899.2024.10651816>. Kuta, Bali, Indonesia.
- [75] Himanshu Gajera, Deepak Nayak, Mukesh Zaveri, A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features, *Biomed. Signal Process Control* 79 (2023) 104186, <https://doi.org/10.1016/j.bspc.2022.104186>.
- [76] M. Tan, Q.V. Le, EfficientNetV2: Smaller Models and Faster Training, 2021, <https://doi.org/10.48550/arXiv.2104.00298> arXiv:2104.00298.
- [77] A. Dosovitskiy, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, <https://doi.org/10.48550/arXiv.2010.11929> arXiv: 2010.11929.
- [78] H. Nassar, D.F. Gleich, Multimodal Network Alignment, 2017, <https://doi.org/10.48550/arXiv.1703.10511> arXiv:1703.10511.

- [79] M. Arnold, D. Singh, M. Laversanne, J. Vignat, S. Vaccarella, F. Meheus, A.E. Cust, E. de Vries, D.C. Whiteman, F. Bray, Global burden of cutaneous melanoma in 2020 and projections to 2040, *JAMA Dermatol.* 158 (5) (2022 May 1) 495–503, <https://doi.org/10.1001/jamadermatol.2022.0160>. PMID: 35353115; PMCID: PMC8968696.
- [80] J.T. Chen, S.J. Kempton, V.K. Rao, The economics of skin cancer: an analysis of medicare payment data, *Plast Reconstr Surg Glob Open* 4 (9) (2016 Sep 27) e868, <https://doi.org/10.1097/GOX.0000000000000826>. PMID: 27757333; PMCID: PMC5054999.
- [81] H. Tsao, M.B. Atkins, A.J. Sober, Management of cutaneous melanoma, *N. Engl. J. Med.* 351 (10) (2004 Sep 2) 998–1012, <https://doi.org/10.1056/NEJMra041245>. Erratum in: *N Engl J Med.* 2004 Dec 2;351(23):2461. PMID: 15342808.
- [82] I. Ajzen, The theory of planned behavior: frequently asked questions, *Human Behav. Emerging Technol.* 2 (4) (2020) 314–324, <https://doi.org/10.1002/hbe2.195>.
- [83] J.W. Brehm, *A Theory of Psychological Reactance*, Academic Press, 1966.
- [84] A. Vance, M. Siponen, S. Pahnla, Motivating is security compliance: insights from habit and protection motivation theory, *Inf. Manag.* 49 (3–4) (2012) 190–198, <https://doi.org/10.1016/j.im.2012.04.002>.
- [85] M. Fishbein, I. Ajzen, *Predicting and Changing Behavior. the Reasoned Action Approach*, Psychology Press, New York, NY, 2010.
- [86] A.S. Ghazali, J. Ham, E. Barakova, P. Markopoulos, The influence of social cues in persuasive social robots on psychological reactance and compliance, *Comput. Hum. Behav.* 87 (2018) 58–65, <https://doi.org/10.1016/j.chb.2018.05.016>.
- [87] Thong Venkatesh, Xu, Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology, *MIS Q.* 36 (1) (2012) 157, <https://doi.org/10.2307/41410412>.
- [88] P. Philip, J.A. Micoulaud-Franchi, P. Sagaspe, E. Sevin, J. Olive, S. Bioulac, A. Sauteraud, Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders, *Sci. Rep.* 7 (2017 Feb 16) 42656, <https://doi.org/10.1038/srep42656>. PMID: 28205601; PMCID: PMC5311989.
- [89] L. Rosenstock, “Historical Origin of the Health Believe Model in Becker,” MH, ed, 1974.

4.2. Adaptive Normalization Layer

The concept of adaptive normalization emerged from the systematic analysis of medical data conducted within the framework of the decision-support component of the Virtual Doctor project. During these investigations, it became evident that the incorporation of adaptive normalization layers into neural network architectures led to measurable improvements in learning stability and predictive accuracy. These preliminary findings provided the motivation to pursue a more comprehensive and in-depth examination of the methodology, both from a theoretical and an applied perspective.

Within this research, the formalism of adaptive normalization layers in the broader context of deep learning is introduced, analyzed, and critically discussed. The study encompasses both rigorous theoretical considerations regarding the statistical and mathematical underpinnings of adaptive normalization, as well as empirical validation on real-world biomedical datasets. Through this dual approach, the work aims to demonstrate not only the methodological soundness of adaptive normalization but also its practical utility in enhancing decision-support systems in medical applications.

4.2.1. Contributions

I was primarily responsible for the conceptual development of the research questions, the design and execution of the analytical framework, and the supervision of the methodological implementation. In addition, I majorly prepared and revised the manuscript for journal submission, ensuring the coherence of the argumentation and the scientific presentation of the results.

The Principal Investigator of the project was Prof. Dr. Dominik Heider, who confirms that the description of my contributions is accurate.

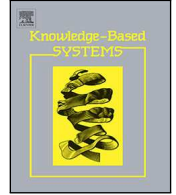
Journal:	Knowledge-Based Systems
Contribution to the publication:	75%
Type of authorship:	First Author
Status of publication:	Published

Münster, 25.11.2025

Place, Date



Prof. Dr. Dominik Heider



Enhancing deep neural network training through learnable adaptive normalization

Jan Benedikt Ruhland ^a, Iraj Masoudian ^a, Dominik Heider ^b

^a Heinrich Heine University Düsseldorf, Faculty of Mathematics and Natural Sciences, Germany

^b University of Münster, Institute of Medical Informatics, Germany

ARTICLE INFO

Dataset link: <https://github.com/JanRland/AdaptiveNormalizationLayer.git>

Keywords:

Artificial intelligence
Machine learning
Deep learning
AI in Medicine

ABSTRACT

Normalization is a fundamental preprocessing technique in data science, commonly used to standardize data distributions prior to model training. Its primary role is to maintain consistent statistical properties across features, which facilitates efficient learning and enhances training stability. In deep learning, normalization methods are particularly beneficial, as they regulate input distributions within neural networks, promoting more stable training and faster convergence.

This study introduces and evaluates a novel approach: learnable adaptive normalization layers integrated into neural networks. Experiments were conducted across nine datasets encompassing feature, image, and time-series data, utilizing various deep learning architectures, including feed-forward, convolutional, and transformer-based neural networks. The results demonstrate that adaptive normalization consistently outperforms traditional methods, such as mean subtraction, standard deviation scaling, and layer normalization. Moreover, in many scenarios, adaptive normalization achieved performance that was either superior to or comparable with batch normalization, with image classification being the notable exception.

These findings indicate that adaptive normalization not only accelerates training convergence but also enhances final model performance in most cases, underscoring its effectiveness. Given that the choice of network architecture is inherently a hyperparameter tuning challenge, we recommend considering adaptive normalization in the preprocessing step for future networks. Furthermore, replacing batch or layer normalization with adaptive normalization may lead to improved training efficiency and final performance, depending on the specific problem.

1. Introduction

Machine learning has become an integral component in critical sectors such as healthcare and cybersecurity [1–6]. Deep learning, a prominent subfield, has shown exceptional capabilities in handling complex classification and regression problems, with successful medical applications, such as cancer diagnostics [7,8] or antimicrobial resistance prediction [9,10]. Recent advancements have focused on building larger models with enhanced performance and developing sophisticated training methodologies [11–13]. A crucial factor influencing these developments is the quality of the training process, which is significantly affected by the preprocessing of input data. One of the primary techniques utilized in this phase is normalization, a method that scales input data to maintain consistency in value ranges and statistical properties. Proper normalization ensures efficient and stable learning dynamics, whereas inadequate normalization can lead to issues like slow convergence, vanishing or exploding gradients, and suboptimal model performance.

Normalization can be implemented using various mathematical techniques. One common approach involves scaling values to a pre-defined range, while another method standardizes data by subtracting the mean and dividing by the standard deviation, which we will call “conventional normalization” throughout this paper. However, challenges arise when dealing with categorical data, such as biological sex or smoking status, which typically cannot be directly normalized using these techniques and often need to be handled separately in the preprocessing pipeline.

Deep neural networks, which consist of multiple layers and a large number of parameters, derive substantial benefits from normalization techniques. Batch normalization, introduced in 2015 by Sergey Ioffe et al. normalizes mini-batches using their mean and variance, while incorporating two additional learnable parameters through an affine transformation [14]. This approach significantly reduces covariate shift during training, leading to faster convergence and reduced need for

* Corresponding authors.

E-mail addresses: jan.ruhland@hhu.de (J.B. Ruhland), irmas100@uni-duesseldorf.de (I. Masoudian), dominik.heider@uni-muenster.de (D. Heider).

<https://doi.org/10.1016/j.knosys.2025.113968>

Received 3 February 2025; Received in revised form 30 April 2025; Accepted 13 June 2025

Available online 5 July 2025

0950-7051/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

extensive hyperparameter tuning, ultimately enhancing the stability and efficiency of the training process.

In 2016, the mathematical framework of batch normalization was extended to standalone normalization layers, retaining the additional two learnable parameters that differentiate it from traditional statistical normalization [15]. In subsequent years, various normalization techniques were introduced to address different use cases. One such method, weight normalization, proposed by Salimans et al. achieved similar benefits to batch normalization by accelerating learning convergence and improving training stability [16].

Instance normalization was developed to mitigate covariate shifts in fast stylization networks, which are generative models used for transferring artistic styles [17]. Unlike batch normalization, which normalizes across height, width, and color channels, instance normalization only operates over spatial dimensions (height and width), earning it the name “contrast normalization” due to its focus on regulating image contrast. In addition to varying the dimensions of normalization, research has also explored new mathematical formulations. One such method, proposed by Chunjie et al. is cosine normalization, which normalizes using the product of the Euclidean norms of the weight and input vectors, providing an alternative approach to standard normalization techniques [18].

The issue of data normalization in deep learning has garnered growing attention in recent years, reflecting its critical importance in the development and performance of machine learning models. An increasing body of literature addresses the challenges associated with normalization techniques, as evidenced by the rising number of publications indexed in the medical research database PubMed under the topic “data normalization challenges”. This trend is visualized in figure A.1 of the supplementary materials, which highlights the escalating interest in this subject within the medical research community. Importantly, the significance of normalization extends beyond the medical domain. It also plays a crucial role in various other areas of artificial intelligence, such as domain adaptation networks [19], where aligning data distributions between source and target domains is essential for achieving robust and generalizable model performance.

This paper builds upon the formalism of batch normalization by introducing a learnable normalization layer, referred to as “adaptive normalization”. This approach eliminates the need for separate normalization during data preprocessing, integrating it directly into the deep learning process. To assess the effectiveness of adaptive normalization, we compared it against batch, layer, and conventional normalization. Experiments were conducted by training deep neural networks on nine datasets and evaluating performance differences on the corresponding test sets. We further expand our study to assess the impact of adaptive normalization on transformer architectures, first introduced in 2017, which have since enabled the development of influential natural language processing (NLP) models such as GPT [20] and Llama [21].

2. Material and methods

2.1. Adaptive normalization layer networks

Our adaptive normalization layer is derived from batch normalization. Rather than computing the mean and variance of incoming data, we utilize learnable parameters $M_{j'k'l'}$ and $V_{j'k'l'}$, which are each the product of two weights $w_{j'k'l'}^0, w_{j'k'l'}^1$. The weight $w_{j'k'l'}^0$ is initialized to the mean of the dataset, while $w_{j'k'l'}^1$ is initially set to 1. This approach facilitates the adjustment of the mean and variance for larger values, as the initial mean may be too large to effectively modify through gradient descent alone. The second weight plays a crucial role in enabling this adjustment. Additionally, we apply the Relu function for the variance to enforce positive values and hence the minimal variance is given by an ϵ constant. The complete affine adjustment is given by Eq. (1).

$$Y_{ijkl} = \frac{X_{ijkl} - M_{j'k'l'}}{\sqrt{V_{j'k'l'} + \epsilon}} \cdot \gamma_{j'k'l'} + \beta_{j'k'l'} \quad (1)$$

$$M_{j'k'l'} = w_{j'k'l'}^0 \cdot w_{j'k'l'}^1$$

$$V_{j'k'l'} = \max \left\{ 0, w_{j'k'l'}^0 \cdot w_{j'k'l'}^1 \right\}$$

The index i denotes the i th sample, while the indices j, k , and l correspond to the dimensions, with j representing the filter, k indicating the x-dimension, and l representing the y-dimension of an image. Additionally, the indices can be set to 0 to signify the absence of a dimension. For data with only one-dimensional features, the equation simplifies by setting k and l to zero. This design allows the integration of the layers into feature, time series, and image data without requiring significant adjustments to either the data or the layer itself. To fasten training, we always initialize the first normalization layer by mean and variance of the respective dataset. We integrate the initialized layer at the commencement of deep neural networks and uninitialized adaptive normalization layers between blocks of linear or convolutional layers to enhance model performance.

In this study, we perform a comparative analysis of our adaptive normalization technique against standard normalization methods, including traditional data normalization, batch normalization, and layer normalization within neural networks. The comparison explores the effects of applying normalization within the network, as opposed to performing it as a preprocessing step. Additionally, we evaluate whether incorporating adaptive normalization layers improves performance relative to batch and layer normalization. The types of comparisons conducted are illustrated on the right side of Fig. 1.

This analysis aims to evaluate the effectiveness and robustness of the proposed adaptive normalization approach across different scenarios. Our adaptive normalization layer functions as a replacement for normalization during both the preprocessing and training phases of the model. However, it is important to note that this integration introduces additional learnable parameters into the network, potentially affecting computational efficiency and increasing model complexity.

2.2. Technical considerations

We identified two key reasons why adaptive normalization is superior to conventional normalization. First, it can prescale categorical data, which is highly prevalent in the medical field and would typically need to be excluded with conventional normalization methods. Second, adaptive normalization addresses the inherent bias present in most datasets more effectively. Unlike conventional normalization, which scales data based solely on existing values, adaptive normalization offers greater flexibility to adjust for biased statistics. To illustrate the effect, we constructed a simple example with a structural causal model. Let X and Y be two normal distributed random variables. The binary target variable is $Z \in \{0, 1\}$ and the probability of Z is fully determined by X and Y .

$$X \sim \mathcal{N}(\mu_x, \sigma_x)$$

$$Y \sim \mathcal{N}(\mu_y, \sigma_y)$$

$$\mathcal{P}(Z = 1|x, y) \propto e^{-\frac{(x+y)^2}{50}}$$

We first designed a neural network with four linear layers, each followed by ReLU activation functions, except for the final layer, which used a sigmoid activation. This network comprised a total of 437 parameters and was always used together with conventional normalization. Next, we developed a second neural network with three linear layers and an adaptive normalization layer at the beginning, reducing the total number of parameters to 177. We then trained the networks on different simulation approaches using stochastic gradient descent with a learning rate of 0.002. Initially, we used the structural causal model with random parameters. Despite having fewer parameters, the network with the adaptive normalization layer demonstrated better performance and faster learning convergence, as illustrated by the training curve 2(a) and the validation curve 2(b). To increase complexity, we modified the input X to consist of three normally distributed

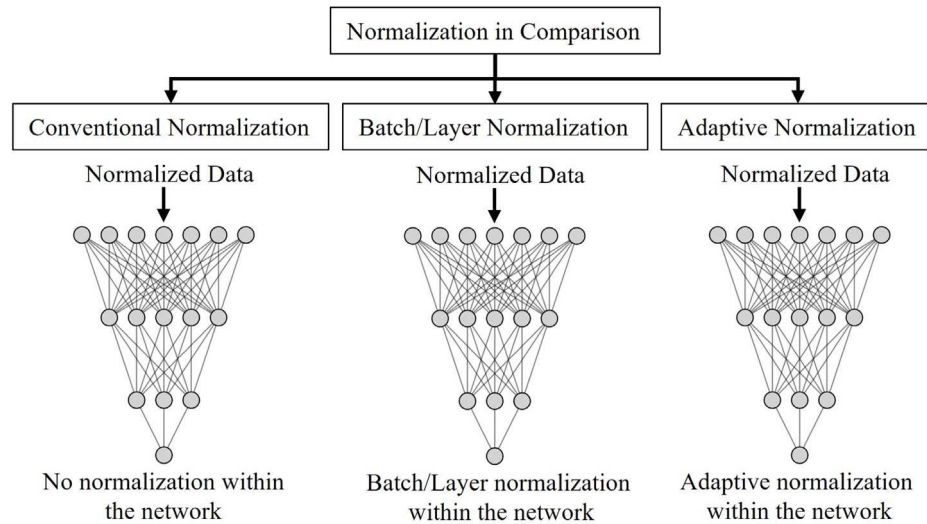


Fig. 1. The image illustrates the comparison between typical normalization procedures. We compare our adaptive normalization with conventional normalization in the preprocessing step and with batch and layer normalization within the networks.

Table 1

The table presents the parameters used in the simulation. In Simulation 1, we utilized only two random variables, whereas in Simulation 2, we defined X to be a combination of three distinct random variables.

Simulation	(μ_x, σ_x)	(μ_y, σ_y)	(μ_1, σ_1)	(μ_2, σ_2)	(μ_3, σ_3)
1	$(-30, 8)$	$(30, 8)$	–	–	–
2	–	$(30, 8)$	$(-50, 6)$	$(-38, 8)$	$(-100, 2)$

random variables ($X \sim X_1 \cup X_2 \cup X_3$), simulating data drawn from different distributions. The input Y was scaled to $2Y$ in order to match the increased value range of X . This setup resembles scenarios where data sources vary, such as a medical study conducted at a pediatric clinic versus one at a general hospital, resulting in biased age distributions.

Such biases may also stem from inherent demographic characteristics of the study location. Additional examples of studies with similarly biased structures are provided in section A of the supplementary materials [22,23]. Although the underlying causal structural model remains unchanged, conventional normalization methods may lead to unstable learning behavior. The results of these simulations are shown in Fig. 2(c) and 2(d). For both simulations, we generated a stratified dataset with 18,000 data points and a validation set of 2000 data points (see Table 1).

Batch normalization was introduced to address internal covariate shift, which arises when parameter updates during gradient descent alter the distribution of activation functions. To stabilize the training process, it normalizes batches by scaling them to have a unit distribution. Similarly, layer normalization operates on the same principle. Adaptive normalization, while also aiming to normalize data, primarily focuses on reducing inherent biases in sub-networks rather than simply preventing unstable training behavior. However, adaptive normalization has a significant drawback: it can destabilize the learning process if there is a large mismatch between the underlying statistical moments of the data and the initialized moments within the adaptive normalization layer. In such cases, gradient descent may fail to adjust the statistical moments in the adaptive normalization layer to appropriate values, impeding effective training.

Adaptive normalization shares many similarities with layer normalization and batch normalization, but differs by offering greater flexibility in adjusting statistical moments. As shown in equations (B1–B5), this flexibility results in a distinct gradient behavior. This also

marks a key difference compared to cosine normalization. While cosine normalization jointly normalizes both the inputs and the weights, it does not alter the number of parameters. Instead, it incorporates the statistical distribution of the layer, thereby exhibiting a closer resemblance to layer normalization [18]. With the introduction of new parameters in the normalization layer, the updates for these parameters are performed using gradient descent. These updates rely on backpropagation, which utilizes the chain rule of differentiation. Consequently, we need to compute the derivative of the adaptive normalization layer Y_{ijkl} with respect to the learnable parameters $w_{j'k'l'}$. This process increases both memory consumption and execution time. In order to make a quantitative assessment we used the designed neural network with four linear layers and a total of 437 parameters and iteratively added adaptive normalization layers. We measured the mean execution time of a training cycle over 10,000 epochs on the same system (Intel Core i7-7700 CPU, 32 GB RAM, NVIDIA GeForce GTX 1060 6 GB, Windows 10) for all six configurations and also listed the number of parameters. The results are shown in Table 2.

2.3. Datasets

We selected datasets to evaluate adaptive normalization across three primary data categories: feature, time series, and image data. The datasets were partially sourced from benchmark datasets and partially from medical studies to emphasize the applicability of adaptive normalization in real-world scenarios.

We utilized three datasets containing only feature data: breast cancer, heart failure prediction, and differentiated thyroid cancer recurrence. The breast cancer dataset, sourced from Kaggle, was collected from the University of Calabar Teaching Hospital's cancer registry and includes 213 patient records with 10 features [24]. Similarly, the heart failure prediction dataset, also available on Kaggle, contains 918 patient records with 12 features, compiled from multiple institutions, including the Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel, V.A. Medical Center Long Beach, and the Cleveland Clinic Foundation [25]. Lastly, the differentiated thyroid cancer recurrence dataset, obtained from the UCI repository, includes 316 patient records with 16 features [26].

The MNIST dataset comprises 60,000 training samples of handwritten digits, with each image represented in a $28 \times 28 \times 1$ format and gray scale intensity values normalized between 0 and 1 [27]. Due to its

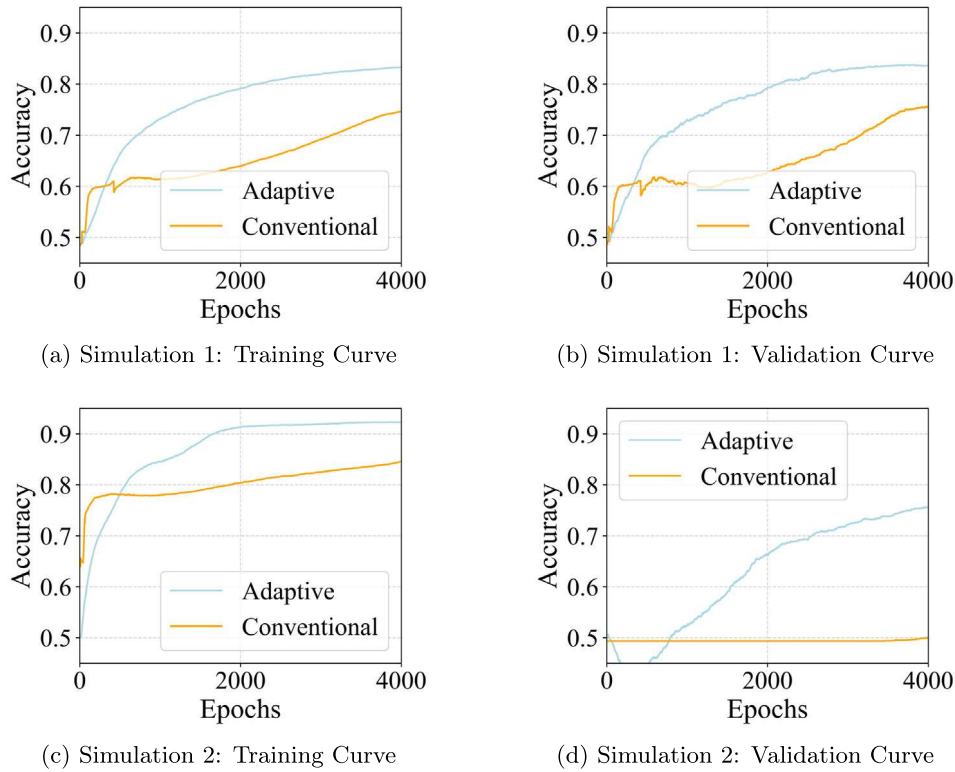


Fig. 2. The figures illustrate the learning curves for two simulation examples of the structural causal model. In Simulation 1, both the adaptive normalization network and the conventionally normalized network achieve similar performance. In Simulation 2, the training curves indicate that both normalization methods again reach comparable performance. However, the validation curve reveals that the network with conventional normalization exhibits signs of overfitting the data.

Table 2

The table presents the execution time and number of parameters of the network for training on the simulated data, with the network gradually incorporating normalization layers.

Layers	Layer type	Execution time [μ s]	Parameters
0	Adaptive Norm	3.49 ± 0.72	437
	Batch Norm	3.49 ± 0.72	437
	Layer Norm	3.49 ± 0.72	437
1	Adaptive Norm	5.86 ± 0.93	449
	Batch Norm	3.80 ± 0.70	441
	Layer Norm	4.57 ± 0.71	441
2	Adaptive Norm	8.33 ± 0.92	473
	Batch Norm	4.40 ± 0.71	449
	Layer Norm	5.65 ± 0.86	449
3	Adaptive Norm	10.64 ± 1.11	569
	Batch Norm	4.87 ± 0.73	481
	Layer Norm	6.68 ± 0.98	481
4	Adaptive Norm	10.68 ± 1.00	665
	Batch Norm	5.41 ± 0.68	513
	Layer Norm	7.56 ± 1.15	513
5	Adaptive Norm	13.69 ± 1.84	689
	Batch Norm	6.08 ± 0.71	521
	Layer Norm	8.49 ± 1.00	521

open-access nature and the relatively straightforward classification task it presents, MNIST is commonly employed for benchmarking neural network architectures. In this study, we utilize the standard training and testing splits provided by the torchvision library [28] to ensure reproducibility and consistency of results.

Given that handwritten digit classification presents a relatively low level of complexity, we incorporated the CIFAR-10 dataset [29] to further evaluate model performance on a more challenging classification task. The CIFAR-10 dataset contains 50,000 color images, each with dimensions $32 \times 32 \times 3$, distributed across 10 diverse classes representing

real-world objects (such as animals and vehicles). These images provide a more complex visual diversity and feature set compared to simple handwritten digits, allowing for a more comprehensive assessment of classification capability. For consistency and reproducibility, we utilized the predefined train-test split included in the CIFAR-10 dataset.

The MIMIC-III dataset comprises pulse recordings from thousands of patients collected from various hospitals [30]. The data was recorded at a sampling rate of 125 Hz using photoplethysmography, a non-invasive optical technique that measures blood volume changes in microvascular tissue. For our analysis, we selected 5000 samples, each lasting 2.1 s, from preprocessed data provided by [31].

In addition to the MIMIC-III dataset, we trained an autoencoder on samples from the PPG-BP database [23] and the AEyEDB pulse dataset. The PPG-BP dataset contains 657 pulse recordings, each 2.1 s long, obtained from 219 patients at a higher sampling rate of 1 kHz using a custom-built device [23]. Furthermore, we incorporated recordings from a study conducted at Heinrich Heine University Düsseldorf, which provided an additional set of data [32]. This dataset consists of 102 recordings, each lasting 60 s, taken from 102 individuals. The data was collected using a commercially available Beurer Pulsoximeter PO 80. To ensure consistency across all datasets, these longer recordings were truncated to produce 612 samples of 2.1 s each. To further standardize our analysis, all samples from the various datasets were resampled to a uniform sampling rate of 100 Hz.

The neural network used for pulse autoencoders and the MNIST classification task is composed of convolutional layers. To further investigate the effectiveness of adaptive normalization within a transformer architecture, we employ electrocardiogram (ECG) data sourced from the MIT-BIH Arrhythmia Database [33]. Preprocessed data from [34] is utilized, containing signal recordings resampled at a 125 Hz sampling rate and clipped into the range between 0 and 1. Each ECG sample is labeled with one of five categories, classifying the underlying cardiac arrhythmia condition. The preprocessed training dataset comprises a total of 87,554 samples, though these are highly imbalanced

across the five arrhythmia categories. Additionally, a final test set is provided, containing 21,892 samples, to evaluate model performance comprehensively.

2.4. Training

We developed three models to evaluate the effectiveness of different normalization techniques, using an 80% train, 10% validation, and 10% test split for the heart disease, differentiated thyroid cancer recurrence, and breast cancer datasets. These models comprised one utilizing adaptive normalization, one employing batch normalization, one incorporating layer normalization, and one based on conventional normalization. All models were trained using binary cross-entropy (BCE) loss in conjunction with the Adam optimizer, an adaptive variant of stochastic gradient descent that incorporates momentum and individual learning rates for each parameter [35].

After training the networks, we applied 10-fold cross-validation on the full dataset and used a paired t-test to evaluate the significance of the differences between models. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was employed as a quantitative metric to assess the discriminative performance of the networks. For the heart disease dataset, the models were trained for 75 epochs with a batch size of 8. Notably, the models utilizing batch normalization and layer normalization operated on conventionally normalized data, whereas the adaptive normalization model processed the raw, unnormalized input. To mitigate class imbalance, we employed Gaussian Noise Upsampling (GNUS), a technique demonstrated to be particularly effective for medical datasets [36]. The learning rate used for this dataset was 0.00022. For the differentiated thyroid cancer recurrence dataset, training was conducted for 90 epochs with a batch size of 10, adhering to the same specifications as the heart disease dataset. The learning rate used for this dataset was 0.000044.

For the breast cancer dataset, the models were trained for 95 epochs with a batch size of 4, once again utilizing the same four models to ensure a uniform training framework. The learning rate used for the breast cancer dataset was 0.00006. This consistent approach across all datasets facilitates a comprehensive evaluation of the normalization techniques' effectiveness in improving model performance.

The MNIST dataset was also divided into 80% for training, 10% for validation, and 10% for testing. The optimal epoch was selected based on performance on the validation set, and the best network weights were saved. These saved weights were then used to evaluate performance on the test set. Training was conducted for 100 epochs using adam optimization [35], with cross entropy as the loss function and a batch size of 50. The learning rate was set to 0.000001.

We trained a small ResNet architecture [37] on the CIFAR-10 dataset. We used 10% of the training set for validation and the predefined test set for final evaluation. The optimization procedure consisted of a batch size of 20, 20 epochs using adam optimization with an initial learning rate of 0.0002 and a weight decay of 0.00001. The learning rate was further reduced by a factor of 0.9 when within 7 epochs no improvement of at least 0.1 was observed.

Similar to the approach used with the MNIST dataset, each pulse dataset was again split into 80% for training, 10% for validation, and 10% for testing. The pulse shape autoencoder was trained for 200 epochs using stochastic gradient descent (SGD), with mean absolute error (MAE) serving as the loss function and a batch size of 20. Each waveform was normalized to the [0, 1] range by subtracting the minimum value of the pulse and dividing by the difference between the maximum and minimum values. The autoencoder architecture was trained independently for each dataset. To compare the performance across different normalization layers, we selected the best-performing network and systematically varied the latent space dimension. This approach allowed us to evaluate the impact of latent space size on network performance and compare it to the other performances. The

MAE across all three datasets was used as a heuristic to determine the optimal latent space dimension for the 2.1-second samples.

For the ECG data, we first quantized the signal values into 401 discrete steps within a range of 0 to 1. This quantization emulates the tokenization process commonly used in NLP tasks. Subsequently, we designed a transformer encoder architecture based on [38] with a vocabulary size of 401, an embedding dimension of 128, and four heads in the multi-head attention layers. We utilized a normed embedding, restricting our comparison of normalization types to those within transformer networks, thereby excluding conventional normalization. Training was performed using the Adam optimizer with an initial learning rate of 0.0001, a batch size of 10, and for a total of 100 epochs. After 10 epochs, the learning rate was steadily reduced by 0.9 to support finer convergence. For each training iteration, we constructed a random, stratified subset of 5000 samples to address class imbalance while maintaining diversity in the training process.

For time series and image data, we applied adaptive normalization following the approach used in layer normalization and also implemented it as contrast normalization. Details of all neural network architectures used in this study are provided in section C of the supplementary materials.

3. Results

Across the breast cancer, thyroid cancer, and heart disease datasets, adaptive normalization consistently achieved the highest overall performance on both the test set and during 10-fold cross-validation. The cross-validation results are presented in Fig. 3. The statistical analysis using the t-test showed significant differences ($p < 0.05$) in nearly all model comparisons, except for the heart disease dataset, where the p -value for the comparison between adaptive and batch normalization was 0.0644. The exact accuracy values for comparing the three different approaches are presented in Table 3, which also includes the classification accuracy for the MNIST, CIFAR-10, and ECG datasets. In contrast to the results from the feature datasets, the training of the MNIST and CIFAR-10 datasets revealed batch normalization to be the best approach followed by adaptive normalization.

A significant performance difference was confirmed only for the CIFAR-10 dataset ($p = 0.019$), while no significant difference was found in the MNIST classification task. The Fig. 5(a) and 5(b) presents the ECG training curves, highlighting the distinct convergence behaviors across normalization techniques. Additionally, 10-fold cross-validation revealed statistically significant differences between adaptive normalization and both batch normalization ($p = 0.039$) and layer normalization ($p = 0.039$). The autoencoder learning task was formulated as a regression task, with the mean absolute error (MAE) values on the test set for autoencoders with a latent dimension of 15 presented in Table 4. Networks utilizing adaptive normalization exhibited significant performance improvements over those using conventional ($p = 0.002$), batch ($p = 0.012$), and layer normalization ($p = 0.012$) techniques. To provide a more comprehensive comparison, we varied the latent space dimension for the adaptive normalization networks across each dataset, as shown in table D.1 of the supplementary materials. This allowed us to compare the performance of the less effective networks to that of the adaptive normalization networks with reduced latent space dimensions. Additionally, the training curves for the autoencoders, displayed in the Figs. 4(a), 4(b), 4(c), and 4(d), illustrate smoother convergence for the adaptive normalization networks, further highlighting their superior training dynamics.

4. Discussion

The performance evaluation of various normalization techniques using the heart disease, breast cancer, and thyroid datasets demonstrated that adaptive normalization consistently ranked among the top-performing methods. In the case of heart disease dataset, adaptive normalization achieved comparable performance to batch normalization (no significant difference).

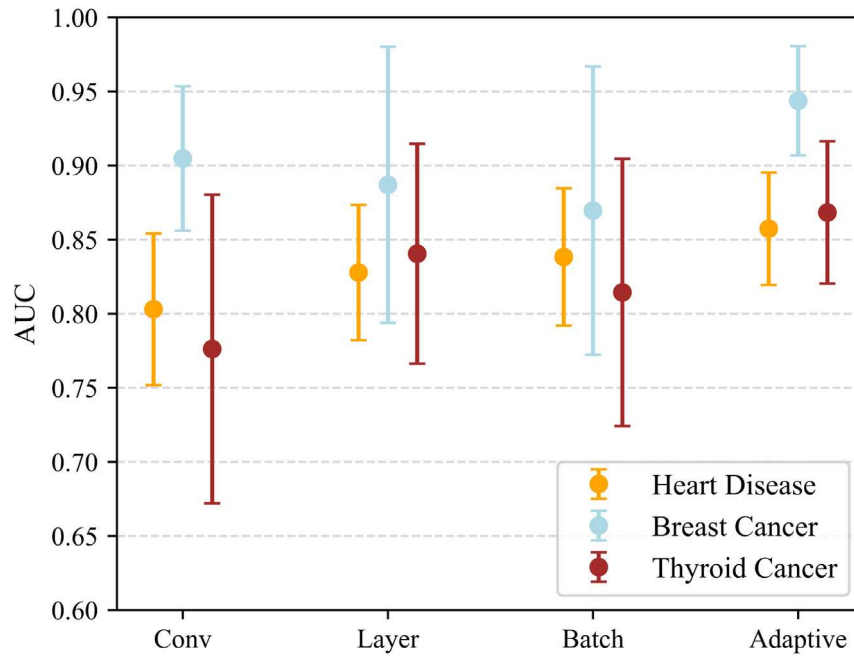


Fig. 3. The figure shows the mean AUC scores for the different normalization techniques for the heart disease dataset, the breast cancer dataset, and the thyroid cancer dataset. The uncertainty is given by the standard deviation of the 10-fold cross-validation. The t-test was used to evaluate the significance of the differences between the models.

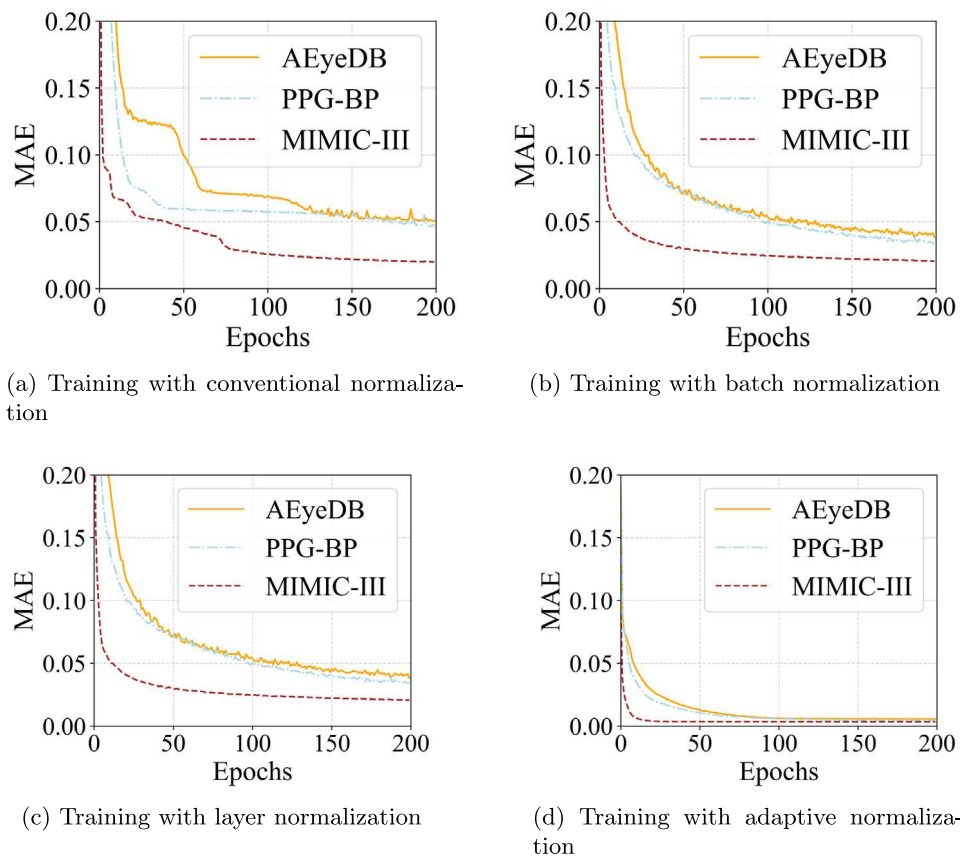


Fig. 4. The figure shows the learning curves on the pulse data training sets for (a) conventional normalization, (b) batch normalization, (c) layer normalization, and (d) adaptive normalization.

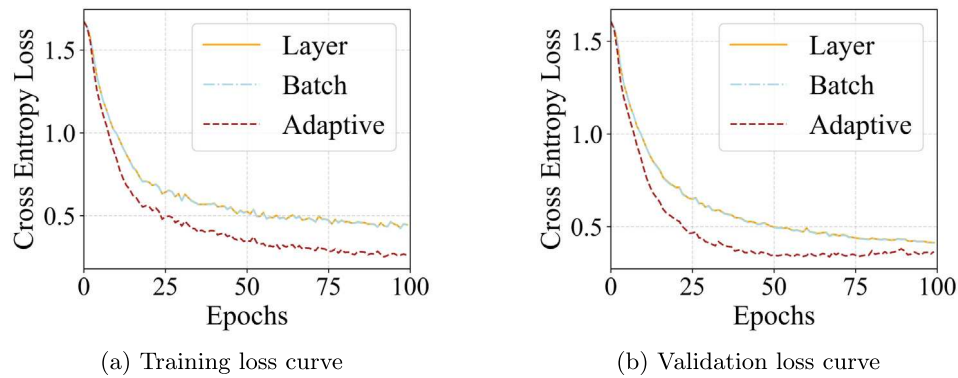


Fig. 5. The figure presents the learning curves for the ECG dataset, comparing three transformer architectures: one incorporating batch normalization, another with layer normalization, and a third with adaptive normalization applied to sequence and embedding dimensions. Subfigure (a) displays the cross-entropy loss on the training set, while subfigure (b) shows the cross-entropy loss on the validation set.

Table 3

The table presents the accuracy scores on the test set for the breast cancer, heart disease, thyroid cancer, MNIST, CIFAR-10, and ECG data set using conventional normalization, batch normalization, layer normalization, and adaptive normalization methods.

Dataset	Conventional	Batch	Layer	Adaptive
Breast Cancer	0.905	0.857	0.905	0.952
Thyroid Cancer	0.763	0.737	0.789	0.816
Heart Disease	0.692	0.736	0.725	0.780
MNIST	0.972	0.979	0.968	0.975
CIFAR-10	0.777	0.821	0.798	0.797
ECG Transformer	–	0.762	0.763	0.825

Table 4

The table presents the MAE scores on the test set for all three pulse datasets using conventional normalization, batch normalization, layer normalization, and adaptive normalization methods.

Dataset	Conventional	Batch	Layer	Adaptive
PPG-BP	0.0434	0.0308	0.0438	0.0042
Mimic-III	0.0184	0.0176	0.0216	0.0034
AEyeDB	0.0657	0.0373	0.0568	0.0069

Our results indicate that adaptive normalization surpasses conventional normalization techniques only in specific image classification tasks. On the MNIST dataset, adaptive normalization outperformed layer normalization, while on the CIFAR-10 dataset, layer normalization achieved higher accuracy. Despite these differences, batch normalization consistently and significantly emerged as the most effective method. This finding indicates that two-dimensional convolutional neural networks gain greater advantages from batch-level normalization, as it helps stabilize high statistical variations in sub-network distributions. This suggests that adaptive normalization cannot adjust quickly enough to sporadic changes. While fine-tuning the initial statistics of adaptive normalization layers might enhance performance, doing so in this comparative analysis would introduce bias, affecting the interpretability and applicability of adaptive normalization.

For instance, if we initialize each adaptive layer (rather than just the first) using the sublayer statistics, it results in no significant difference compared to batch normalization.

The learning curves for the autoencoder trained on pulse datasets exhibit more stable convergence patterns when employing adaptive normalization. This stabilization potentially contributes to enhanced overall performance. When comparing results across different normalization techniques, the discrepancy is substantial. Autoencoders with a latent dimensionality of 15 using conventional normalization methods achieve similar performance to those with a latent dimensionality of approximately 5 when utilizing adaptive normalization. This performance difference may be attributed to the interplay between the

smoother training dynamics introduced by adaptive normalization and the increased capacity provided by the higher number of learnable parameters. Similarly, the transformer architecture for ECG datasets achieved its best performance with adaptive normalization. This parallel could be attributed to the shared fundamental challenges associated with time series data.

It is important to note that while the pulse datasets were resampled to ensure consistency in the analysis, the same convergence patterns are observed even without resampling. Moreover, combining adaptive normalization with weight regularization requires careful consideration. Large initial weights in adaptive normalization can disproportionately influence the gradients, potentially leading to instability during training. To address this issue, the initial weight should be frozen and hence excluded from updates during the training process.

Another potential issue with adaptive normalization stems from the ReLU-based variance estimation, which can introduce discontinuities in the gradient. One possible mitigation strategy is to replace the ReLU function with a Gaussian Error Linear Unit (GELU) [39]. However, to properly adapt Eq. (1), the minimum value of the GELU function must be incorporated into the ϵ term. Without this adjustment, singularities or complex-valued results may occur. In real-world applications, particularly in the medical field, hardware requirements and execution time are critical factors and are often closely linked. While adaptive normalization increases the computational complexity of the gradient descent optimization algorithm by a factor of six (see section B of the supplementary materials), it can also lead to significant performance improvements. This suggests that smaller networks incorporating adaptive normalization may outperform larger networks without it. Since a common and practical strategy for developing optimal network architectures is to begin with smaller models, adaptive normalization presents itself as a promising method to consider during the design process. While our work has primarily focused on medical applications of adaptive normalization, the challenge of acquiring data without statistical shifts is prevalent across all domains of applied machine learning. Several alternative methods have been proposed to address issues caused by distributional shifts, such as the ask-oriented theil index-based meta-learning network [40], the digital twin-assisted multiscale feature fusion network [41], and the dynamic collaborative adversarial domain adaptation network [19]. These approaches are particularly relevant in areas like fault diagnosis and engineering, where data variability can significantly impact model performance. The existence and continued development of such methods highlight the broader relevance and importance of the adaptive normalization research.

5. Conclusion

Data normalization is an essential part of the preprocessing pipeline in machine learning models. Here, we present adaptive normalization, a

learnable technique inspired by batch normalization. Unlike traditional normalization methods, it integrates directly into deep learning models, eliminating the need for manual storage of normalization parameters, a clear advantage for machine learning applications. Additionally, it can handle various data types. Our results show that adaptive normalization consistently matches or outperforms traditional, layer, and batch normalization techniques for feature and time series data. For image data, however, batch normalization remains the best approach.

Our experiments, conducted using medical datasets such as heart disease, breast cancer, and thyroid data, demonstrate that adaptive normalization leads to faster convergence and superior performance compared to other normalization methods. This observation underscores the significance of adaptive normalization, particularly for models dealing with medical data, where achieving high accuracy and stability is paramount.

Our research further suggests that the advantages of adaptive normalization become increasingly evident in time series data. This is exemplified by experiments using pulse data with autoencoders, where adaptive normalization provided notable performance gains. A similar trend was observed when comparing results from the ECG dataset, indicating that adaptive normalization may be particularly beneficial for time series data.

We did not observe any performance improvement when comparing adaptive normalization to batch normalization in image classification tasks. This could be attributed to the high mean and variance values calculated within the network during normalization. Similarly, when initializing the first adaptive normalization layer, the method is subjected to gradient descent and scaled updates. These processes might hinder the layer from achieving sufficiently high mean or variance values, potentially limiting its effectiveness.

Our research indicates that incorporating adaptive normalization along the embedding and sequence dimensions can be advantageous for prominent transformer architectures. This highlights the potential impact of adaptive normalization in advancing natural language processing methodologies and emphasizes its significance in this field.

The selection of an appropriate normalization method is framed as a hyperparameter optimization problem. Our analysis demonstrates that different normalization techniques influence both the training dynamics and the final model performance. The results suggest that adaptive normalization is a robust and effective choice across various analytical scenarios.

In future work, we aim to explore the integration of adaptive normalization into transfer learning, adaptive data augmentation, and domain adaptation scenarios. Given that adaptive normalization dynamically adjusts to input data distributions, it may be possible to facilitate knowledge transfer by retraining only the adaptive normalization layers while keeping the rest of the network parameters fixed. This approach could enable efficient adaptation to new domains with minimal retraining. A similar strategy could be employed for domain adaptation, where adjusting only the normalization layers may suffice to account for distributional shifts across domains. Additionally, since certain data augmentation techniques modify the underlying data distribution, we plan to investigate how such changes interact with adaptive normalization and whether it can maintain performance robustness under these transformations.

CRedit authorship contribution statement

Jan Benedikt Ruhland: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Iraj Masoudian:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Dominik Heider:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.knosys.2025.113968>.

Data availability

An implementation of the adaptive normalization layer can be found on <https://github.com/JanRland/AdaptiveNormalizationLayer.git>.

References

- [1] A. Handa, A. Sharma, S. Shukla, Machine learning in cybersecurity: A review, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 9 (2019) e1306.
- [2] S. Mahdaviifar, A. Ghorbani, Application of deep learning to cybersecurity: A survey, *Neurocomputing* 347 (2019).
- [3] F. Ullah, H. Naem, S. Jabbar, S. Khalid, M.A. Latif, F. Al-turjman, L. Mostarda, Cyber security threats detection in internet of things using deep learning approach, *IEEE Access* 7 (2019) 124379–124389.
- [4] M.A. Ferrag, L. Maglaras, S. Moschogiannis, H. Janicke, Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study, *J. Inf. Secur. Appl.* 50 (2020) 102419.
- [5] H. Habebh, S. Gohel, Machine learning in healthcare, *Curr. Genomics* 22 (2021).
- [6] R. Miotto, F. Wang, S. Wang, X. Jiang, Deep learning for healthcare: review, opportunities and challenges, *Brief. Bioinform.* 19 (2017).
- [7] M. Dildar, S. Akram, M. Irfan, H.U. Khan, M. Ramzan, A.R. Mahmood, S.A. Alsaiani, A.H.M. Saeed, M.O. Alraddadi, M.H. Mahnashi, Skin cancer detection: A review using deep learning techniques, *Int. J. Environ. Res. Public Heal.* 18 (10) (2021) 5479.
- [8] P. Danaee, R. Ghaeini, D. Hendrix, A deep learning approach for cancer detection and relevant gene identification, *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 22 (2017) 219–229.
- [9] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, O. Schwengers, D. Heider, Deep transfer learning enables robust prediction of antimicrobial resistance for novel antibiotics, *Antibiotics* 11 (11) (2022) 1611.
- [10] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, O. Schwengers, D. Heider, Multi-label classification for multi-drug resistance prediction of *Escherichia coli*, *Comput. Struct. Biotechnol. J.* 20 (2022) 1264–1270, URL <http://dx.doi.org/10.1016/j.csbj.2022.03.007>.
- [11] R. Gozalo-Brizuela, E.C. Garrido-Merchan, ChatGPT is not all you need. a state of the art review of large generative AI models, 2023, *ArXiv E-Prints*.
- [12] A. Baevski, H. Zhou, A. Mohamed, M. Auli, Wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020, *ArXiv E-Prints*.
- [13] O.A. Sarumi, D. Heider, Large language models and their applications in bioinformatics, *Comput. Struct. Biotechnol. J.* 23 (2024) 3498–3505.
- [14] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, *ArXiv E-Prints*.
- [15] J. Lei Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, *ArXiv E-Prints*.
- [16] T. Salimans, D.P. Kingma, Weight normalization: A simple reparameterization to accelerate training of deep neural networks, 2016, *ArXiv E-Prints*.
- [17] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, 2016, *ArXiv E-Prints*.
- [18] L. Chunjie, Z. Jianfeng, W. lei, Y. Qiang, Cosine normalization: Using cosine similarity instead of dot product in neural networks, 2017.
- [19] X. Wang, H. Jiang, M. Mu, Y. Dong, A dynamic collaborative adversarial domain adaptation network for unsupervised rotating machinery fault diagnosis, *Reliab. Eng. Syst. Saf.* 255 (2025) 110662.
- [20] A. Radford, K. Narasimhan, Improving language understanding by generative pre-training, 2018.
- [21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, 2023, *ArXiv E-Prints*.
- [22] E. Korot, N. Pontikos, X. Liu, S. Wagner, L. Faes, J. Huemer, K. Balaskas, A. Denniston, A. Khawaja, P. Keane, Predicting sex from retinal fundus photographs using automated deep learning, *Sci. Rep.* 11 (2021).
- [23] M. Elgendi, V. Galli, C. Ahmadzadeh, C. Menon, Dataset of psychological scales and physiological signals collected for anxiety assessment using a portable device, *Data* 7 (2022) 132.

- [24] I. Eteng, E. Bisong, T. Fagbola, M. Ibrahim, J. Udosen, S. Akpotuzor, UCTH Breast Cancer Dataset, Mendeley Data, 2023.
- [25] Fedesoriano, Heart failure prediction dataset, 2021.
- [26] S. Borzooei, A. Tarokhian, Differentiated thyroid cancer recurrence, 2023, DOI: <https://doi.org/10.24432/C5632J>, UCI Machine Learning Repository.
- [27] Y. LeCun, C. Cortes, C. Burges, MNIST handwritten digit database, 2, 2010, ATT Labs [Online]. Available: <http://yann.lecun.com/exdb/mnist>.
- [28] TorchVision maintainers and contributors, TorchVision: PyTorch's Computer Vision library, GitHub, 2016, <https://github.com/pytorch/vision>.
- [29] A. Krizhevsky, V. Nair, G. Hinton, CIFAR-10 (Canadian Institute for Advanced Research).
- [30] A. Johnson, T. Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark, MIMIC-III, a freely accessible critical care database, *Sci. Data* 3 (2016) 160035.
- [31] F. Schrumpf, P. Frenzel, C. Aust, G. Osterhoff, M. Fuchs, Assessment of deep learning based blood pressure prediction from ppg and rPPG signals, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3820–3830.
- [32] G.C.T. Register, Machine learning based health evaluation using retina images and recorded pulse waves, 2024, <https://drks.de/search/en/trial/DRKS00033094>, (Accessed 30 November 2024).
- [33] G. Moody, R. Mark, The impact of the MIT-bih arrhythmia database, *IEEE Eng. Med. Biol. Mag.* 20 (3) (2001) 45–50.
- [34] M. Kachuee, S. Fazeli, M. Sarrafzadeh, ECG heartbeat classification: A deep transferable representation, 2018, CoRR, [abs/1805.00794](https://arxiv.org/abs/1805.00794).
- [35] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, *ArXiv E-Prints*.
- [36] J. Beinecke, D. Heider, Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making, *BioData Min.* 14 (49) (2021).
- [37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016, p. 1.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, *ArXiv E-Prints*.
- [39] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), 2023.
- [40] M. Mu, H. Jiang, X. Wang, Y. Dong, A task-oriented theil index-based meta-learning network with gradient calibration strategy for rotating machinery fault diagnosis with limited samples, *Adv. Eng. Inform.* 62 (2024) 102870.
- [41] Y. Dong, H. Jiang, Z. Wu, Q. Yang, Y. Liu, Digital twin-assisted multiscale residual-self-attention feature fusion network for hypersonic flight vehicle fault diagnosis, *Reliab. Eng. Syst. Saf.* 235 (2023) 109253.

4.3. Pathological Anomaly Detection in Fundus Images

Following the construction of the first Virtual Doctor prototype and its installation at the University Hospital Bochum, further opportunities for extending the scope of the clinical decision-support component were explored. One particularly promising direction was the incorporation of ophthalmological diagnostics, as fundus imaging has proven to be a highly valuable modality for the detection of a broad spectrum of ocular diseases.

To investigate this potential, we initiated and conducted a dedicated study at the Heinrich Heine University Düsseldorf. Within the framework of this study, a total of 204 fundus images were collected from 104 participants under carefully controlled conditions. The ethical approval for the study was obtained through the appropriate institutional review process, with the approval letter provided in the appendix. This dataset formed the empirical foundation for the subsequent development of a deep learning-based anomaly detection system, designed to identify and localize pathologically suspicious regions within fundus images. The methodology aimed not only to improve diagnostic accuracy but also to generate interpretable outputs highlighting relevant regions in the images, thereby assisting physicians in their decision-making processes.

4.3.1. Contributions

I was primarily responsible for drafting the documentation required for ethical committee submission, as well as for revising the study protocol in response to feedback. Furthermore, I directly conducted the study, including participant management and the acquisition of fundus images, ensuring both the quality and consistency of the collected data. On the methodological side, I played a leading role in the conceptualization and implementation of the anomaly detection framework, as well as in the subsequent analytical evaluations. Beyond the technical contributions, I assumed the role of first author of the associated manuscript, writing the majority of the text and ensuring the coherent integration of all project components into a unified scholarly presentation.

The Principal Investigator of the project was Prof. Dr. Dominik Heider, who confirms that the description of my contributions is accurate.

Journal:	BioData Mining
Contribution to the publication:	75%
Type of authorship:	First Author
Status of publication:	Submitted
Preprint:	https://arxiv.org/abs/2511.18627

Münster, 25.11.2025

Place, Date



Prof. Dr. Dominik Heider

Functional Localization Enforced Deep Anomaly Detection Using Fundus Images

Jan Benedikt Ruhland¹, Thorsten Papenbrock², Jan-Peter Sowa³, Ali Canbay³, Nicole Eter⁴, Bernd Freisleben², Dominik Heider^{5*}

1: Heinrich Heine University Düsseldorf, Faculty of Mathematics and Natural Sciences, Düsseldorf, Germany

2: Philipps University of Marburg, Department of Mathematics and Computer Science, Marburg, Germany

3: University Hospital Knappschaftskrankenhaus Bochum, Department of Internal Medicine, Germany

4: University of Münster, Department of Ophthalmology, Münster, Germany

5: University of Münster, Institute of Medical Informatics, Münster, Germany

*corresponding author: dominik.heider@uni-muenster.de

Abstract

Reliable detection of retinal diseases from fundus images is challenged by the variability in imaging quality, subtle early-stage manifestations, and domain shift across datasets. In this study, we systematically evaluated a Vision Transformer (ViT) classifier under multiple augmentation and enhancement strategies across several heterogeneous public datasets, as well as the AEyeDB dataset, a high-quality fundus dataset created in-house and made available for the research community. The ViT demonstrated consistently strong performance, with accuracies ranging from 0.789 to 0.843 across datasets and diseases. Diabetic retinopathy and age-related macular degeneration were detected reliably, whereas glaucoma remained the most frequently misclassified disease. Geometric and color augmentations provided the most stable improvements, while histogram equalization benefited datasets dominated by structural subtlety. Laplacian enhancement reduced performance across different settings.

On the Papila dataset, the ViT with geometric augmentation achieved an AUC of 0.91, outperforming previously reported convolutional ensemble baselines (AUC of 0.87), underscoring the advantages of transformer architectures and multi-dataset training. To complement the classifier, we developed a GANomaly-based anomaly detector, achieving an AUC of 0.76 while providing inherent reconstruction-based explainability and robust generalization to unseen data. Probabilistic calibration using GUESS enabled threshold-independent decision support for future clinical implementation.

Dataset-level analysis highlighted that acquisition consistency substantially impacts model robustness. Curated datasets with manual quality control yielded the highest performance. Overall, our results support the combined use of transformer-based classifiers and principled anomaly detection for fundus analysis and motivate future work on domain adaptation and improved calibration through larger healthy cohorts.

Keywords: Machine Learning, AI in medicine, Explainable AI, Anomaly Detection, Fundus Images

1. Introduction

Fundus imaging is a non-invasive, cost-effective, and widely accessible modality used for visualizing the retina and its surrounding structures. It plays an essential role in the early detection and monitoring of retinal disorders such as age-related macular degeneration (AMD) [1, 2], diabetic retinopathy (DR) [3], and glaucoma [4]. These diseases exhibit distinct visual patterns, ranging from microaneurysms and hemorrhages to optic disc cupping and macular degeneration, which can be captured in fundus photographs. However, the visibility and distinctiveness of these features vary strongly. While some abnormalities are easily recognizable, even by non-specialists, others manifest as subtle structural or textural deviations that may elude expert ophthalmologists. Nevertheless, the affordability, non-invasive nature, and scalability of fundus imaging makes it indispensable for both clinical diagnosis and large-scale screening programs.

Beyond its clinical significance, fundus imaging has become a central focus in medical image analysis research, driving progress in vessel segmentation, lesion localization, synthetic data generation, and automated disease classification [5, 6, 7, 8]. Recent studies have demonstrated that retinal fundus images can be used not only to diagnose specific ophthalmic diseases but also to infer systemic health conditions such as cardiovascular risk, anemia, and even biological aging [9, 10, 11]. Moreover, datasets such as JSIEC [6] reveal that fundus images can display features corresponding to more than 39 distinct pathological abnormalities, each representing a unique manifestation of retinal or systemic disease. This vast pathological diversity underscores a critical challenge for conventional supervised classification approaches. Training deep neural networks to accurately classify dozens of disease types requires extensive labeled datasets and suffers from severe class imbalance. Furthermore, rare or previously unseen abnormalities may be misclassified, limiting the reliability and generalizability of such systems in real-world clinical settings.

Deep learning models designed for fundus image analysis, including convolutional neural networks (CNNs) and transformer-based architectures, have achieved remarkable accuracy in recent studies [12, 13, 14]. However, their deployment in clinical environments is constrained by several factors. First, these models are computationally demanding, requiring high-end GPUs and large memory footprints, and second, they often depend on cloud-based processing, which raises concerns about patient privacy and data security. Additionally, despite their predictive power, such models generally act as black boxes. They output categorical labels or probabilities without explaining the underlying rationale behind each decision. This lack of interpretability hinders clinical validation and erodes trust among medical practitioners, who must understand and verify the reasoning behind AI-driven assessments before integrating them into patient care.

Recent advances in explainable artificial intelligence (AI) aim to address these challenges by identifying and visualizing the image regions most influential in model predictions. Frameworks such as SHAP [15] and DeepLIFT [16] quantify feature importance, while attention-based and gradient-based visualization methods generate saliency maps or contribution maps highlighting diagnostically relevant areas [17, 18]. By localizing key retinal structures or lesions that contribute to the final decision, these methods enhance transparency, facilitate clinical interpretability, and allow for more informed validation of AI-generated results.

Given the large number of potential retinal pathologies and the limitations of fixed-category classification, anomaly detection offers a more scalable and clinically realistic alternative. Instead of explicitly classifying every possible disease, anomaly detection frameworks learn the normal retinal appearance and identify deviations indicative of a pathology. This paradigm is particularly well-suited for medical imaging, where the spectrum of disease manifestations is vast and often unpredictable, e.g., for rare diseases.

In this study, we present a comprehensive framework for explainable retinal image analysis that integrates transformer-based classification, reconstruction-driven anomaly detection, and probabilistic calibration. Our work systematically evaluates the performance and robustness of a Vision Transformer classifier across multiple augmentation strategies, investigates a constrained attention mechanism for embedded localization, and develops an anomaly detection pipeline based on GANomaly [19] combined with GUESS [20] calibration. In addition, we introduce AEyeDB, a newly curated, high-quality fundus image dataset created by our group and made available to the research community upon request.

The main contributions of this research are:

1. Systematic Evaluation of Transformer-Based Fundus Classification

We provide a detailed assessment of ViT performance under geometric, chromatic, and contrast-enhancing augmentation strategies across multiple heterogeneous datasets. Our results show that while geometric and color augmentations offer stable improvements, histogram equalization and Laplacian enhancement can degrade performance, providing practical guidance for preprocessing choices in fundus classification pipelines.

2. Embedded Attention Mechanism for Inherent Localization

We incorporate a constrained attention-based masking strategy directly into the ViT architecture to promote inherent spatial interpretability. Although this mechanism did not directly improve classification accuracy, it produced consistent and physiologically meaningful attention maps, demonstrating the feasibility of integrating functional localization into end-to-end training without relying on post-hoc saliency methods.

3. Reconstruction-Based Anomaly Detection with Explainability

We extend the GANomaly framework to provide interpretable anomaly maps derived from reconstruction error distributions. Although additional loss terms (mask loss, KL regularization) did not improve performance, the baseline model achieved an AUC of 0.76 and generalized well to unseen data. The reconstruction maps reliably highlighted pathological structures, offering a principled and transparent alternative to black-box classifier outputs.

4. Probabilistic Calibration With GUESS

We employ the GUESS framework to transform raw anomaly scores into calibrated abnormality probabilities. This threshold-independent approach produces well-ranked probabilistic outputs and facilitates clinically meaningful decision support. The use of AUC as

the primary metric aligns with the goal of evaluating probability ordering rather than thresholded classification.

5. Introduction of the AEyeDB Dataset

We release AEyeDB, a curated, high-quality fundus imaging dataset collected and annotated by our team. The dataset will be accessible upon request to researchers who provide a brief description of their intended research aims, promoting transparency, reproducibility, and community collaboration.

Research in Context

The development of multi-head attention in large language models [21] has inspired the adaptation of attention mechanisms to the image domain through architectures such as the Vision Transformer [22]. ViTs leverage self-attention to assign varying importance to different spatial regions of an image, allowing the model to dynamically focus on informative features while ignoring irrelevant areas. This inherent ability to localize salient regions makes attention mechanisms particularly attractive for medical imaging, where understanding why a model predicts a certain outcome is as important as the prediction itself.

Attention mechanisms have been widely applied in medical imaging tasks, including fundus disease classification [6, 23], chest X-ray interpretation [24], and skin lesion detection [25]. In these applications, attention maps provide interpretable visualizations that highlight regions contributing most to a model’s decision, thereby improving clinician trust and enabling better validation of AI-driven diagnoses. In fundus imaging specifically, prior studies have focused primarily on disease classification and post-hoc interpretability. For instance, an ensemble of convolutional neural networks (CNNs) was used to classify 39 fundus pathologies, achieving a weighted F1 score of 92.3% while employing SHAP values to enhance model explainability [6]. Similarly, ViT-based architectures have been employed to generate attention maps for visual explanations in retinal disease classification [26]. These works typically rely on self-attention mechanisms as introduced by Vaswani et al. [21], where the focus of the network is determined dynamically but independently of the anomaly detection objective.

While these methods provide useful post-hoc interpretations, they often require additional processing after training, and their ability to identify subtle or previously unseen pathological features remains limited. Traditional interpretability techniques such as SHAP [15], CAM [27], and Grad-CAM [28] quantify feature contributions after model training, which can create a gap between the learning process and the explanatory outputs. In contrast, our approach integrates an explainable masking mechanism directly into the network architecture, making interpretability an inherent component of model learning. This design allows the model to learn which regions are diagnostically relevant during training, rather than inferring importance post hoc.

Moreover, while most prior studies have focused exclusively on either classification or localization, our framework combines both objectives in a unified architecture. We introduce sparsity through an attention mask that guides the learning process, enforcing the model to concentrate on the most informative regions while suppressing irrelevant features. This attention-guided mechanism is then applied within a discriminative anomaly detection framework, enabling the identification of pathological anomalies even in cases where explicit class labels are unavailable or diseases are rare.

By integrating functional localization with anomaly detection, our method addresses a key limitation of prior work: the difficulty to generalize across a large spectrum of abnormalities, which is critical given that biomarkers like fundus images can contain over 39 distinct pathological patterns.

Finally, our study investigates whether the incorporation of learned attention masks enhances both performance and interpretability of disease detection models. By embedding localization directly within the learning process and coupling it with anomaly detection, we provide a scalable, transparent, and clinically relevant framework for fundus image analysis, distinct from traditional post-hoc attention or classification-only approaches.

2. Materials and Methods

2.1 Proposed Models

The proposed framework builds upon the Vision Transformer architecture and extends it through a constrained attention mechanism for interpretable localization and anomaly detection. As a baseline model, we employ the ViT-B/16 configuration containing approximately 86 million trainable parameters [22]. The pretrained model is first used to evaluate domain shifts and the effect of various data augmentation strategies on fundus images. The most effective configuration is then adapted to train a segmentation network with constrained attention, allowing the model to highlight diagnostically relevant regions during inference.

For spatial localization, we employ a U-Net based segmentation network endowed with adaptive normalization [29] and self-attention [21] in the bottleneck region (described in Appendix A). The network outputs a single-channel mask $M=g_\phi(x)\in[0,1]$ through a sigmoid activation, where each pixel value represents the estimated importance of that region. This mask is applied elementwise to the input image, producing a masked image $x'=M\odot x$, which is subsequently passed into the ViT classifier f_θ that has been pretrained on multi-class fundus classification. The model prediction is expressed as $y'=f_\theta(x')$.

The segmentation network g_ϕ is optimized using a composite loss function that combines the standard cross-entropy loss from the classifier with an L_1 regularization term to encourage sparsity in the generated masks. The total loss is formulated as

$$L=L_{CE}(f_\theta(g_\phi(x)\odot x), y)+\lambda\|M\|_1 \quad (1)$$

where the cross-entropy loss L_{CE} enforces prediction accuracy relative to the ground truth, and the regularization coefficient λ controls the trade-off between classification fidelity and mask sparsity. By minimizing this loss, the network learns to identify minimal yet sufficient regions required for accurate classification, effectively removing irrelevant background and noise. This process serves as an embedded preprocessing step that enhances both interpretability and robustness.

The resulting attention masks can be interpreted as learned importance maps that highlight regions influencing the classifier’s decision, similar in spirit to post-hoc interpretability methods such as SHAP and Grad-CAM [15, 28]. However, unlike these approaches, which operate after model training, our constrained attention mechanism is fully integrated into the training process, making localization an intrinsic part of learning.

Building on this foundation, the attention mechanism is extended to an anomaly detection framework that leverages the generative-adversarial architecture of GANomaly [19]. The model follows an encoder-decoder-encoder structure in which the input image is first encoded into a latent representation, reconstructed by the decoder, and then re-encoded to assess reconstruction consistency. Deviations between original and reconstructed latent features form the basis of the anomaly score $A(x)$. In the classical formulation, GANomaly employs three loss components: an L_2 adversarial feature map loss, an L_1 reconstruction loss, and an L_2 latent consistency loss.

Inspired by recent developments in variational autoencoders [30] and diffusion models [31], we modify this architecture by introducing a Gaussian prior regularization on the latent space through a Kullback-Leibler (KL) divergence term. This additional constraint enforces a structured, stable latent distribution, improving reconstruction fidelity and anomaly discrimination [31]. Two configurations are therefore explored: the vanilla GANomaly model and the KL-regularized variant. The better configuration is then also trained in conjunction with the constrained attention mask introduced earlier, allowing us to evaluate how attention-guided learning influences anomaly detection performance.

The integration of the attention mask into the autoencoder training pipeline further enhances the reconstruction process by emphasizing diagnostically relevant retinal regions. This targeted focus reduces redundancy and prevents background features from diluting the anomaly signal. Consequently, the generative model learns to reconstruct only the most informative areas, improving the spatial alignment between detected anomalies and actual pathological regions in the fundus images.

Technically, the integration is achieved by introducing an additional reconstruction loss term applied selectively to the top 90th percentile of the attention-weighted image intensities. This selective penalization ensures that the model prioritizes reconstruction accuracy in regions deemed clinically significant, while disregarding low-importance background areas.

Finally, the computed anomaly score $A(x)$ is calibrated using the GUESS framework [20]. Unlike conventional GANomaly methods that normalize scores empirically over healthy and abnormal datasets, GUESS provides a probabilistic calibration based on Bayes' theorem, yielding $P(\text{pathology}|A(x))$. This transforms raw anomaly scores into interpretable probabilities of pathology, thereby enhancing clinical interpretability and reliability.

Overall, this integrated approach combines discriminative attention, generative reconstruction, and probabilistic calibration in a unified pipeline. The ViT-based classifier establishes a robust and interpretable baseline, while the constrained attention mask enforces spatial focus. The KL-regularized GANomaly model captures deviations from the learned healthy distribution, and GUESS converts these deviations into probabilistic indicators of disease. Together, these elements form a cohesive system for interpretable, accurate, and generalizable anomaly detection in retinal fundus imaging.

2.2 Experimental Procedure

2.2.1. Fundus Image Datasets

We conducted a study at the Heinrich Heine University Düsseldorf involving 103 young adults (age ≥ 18 , written consent provided). Participants were informed on the potential risks of phototoxicity and epileptic triggering associated with the image recording process. Subsequently, they completed a questionnaire to collect data on age, gender, and preexisting conditions/diseases. Fundus images from

the participants were captured using the Rodenstock FundusScope. The Beurer PO 80 was used to record the pulse waveform of the participants.

The average age in the study was 26.93 ± 7.18 years (see also Figure 1). Out of the 103 participants, 21 were female and 82 were male. In the provided questionnaire, 13 participants reported a pre-existing condition which might influence the retina recording, e.g., diabetes or hypertension. Five participants took part in the study multiple times, resulting in a total of 256 fundus images. For the analysis, we utilized a subset of 204 images to minimize potential biases arising from multiple images of the same individual and to exclude recordings of insufficient quality.

The study was approved by the ethical committee at Heinrich Heine University and registered (<https://drks.de/search/de/trial/DRKS00033094>) to ensure transparency and adherence to ethical research standards.

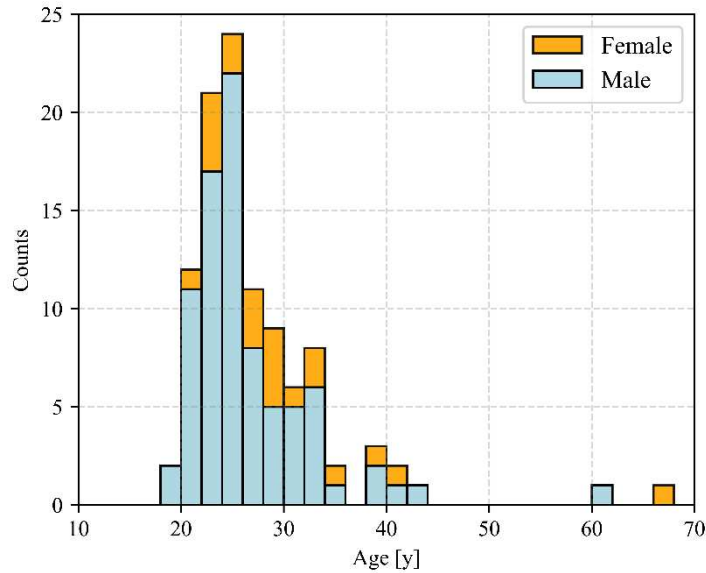


Figure 1: The figure illustrates the age distribution for the AEyEDB and the bin width is two years. Within each bin, the blue area indicates the proportion of male patients, while the orange area represents the fraction of female patients.

In addition to the AEyEDB dataset, several publicly available datasets were employed to enhance model generalizability and enable cross-dataset evaluation. These include FIVES [5], Mendeley [32], Papila [33], Messidor [34], and JSIEC [6, 35].

The FIVES dataset comprises 800 fundus images collected at the Ophthalmology Centre of the Second Affiliated Hospital of Zhejiang University. The images are evenly distributed across four diagnostic categories: healthy, diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD) (200 images per class). The dataset also provides image quality annotations and was recorded using a Topcon TRC-NW8 fundus camera under uniform imaging conditions. The consistent acquisition protocol makes FIVES a suitable external reference dataset for evaluating cross-domain performance.

The Mendeley dataset is a large compilation of publicly available retinal fundus images sourced from multiple clinical studies. It contains diverse pathological categories including diabetic retinopathy,

glaucoma, macular scar, retinitis pigmentosa, and disc edema, with over 5,000 high-resolution images spanning healthy and diseased retinas. All images are labeled according to the predominant clinical diagnosis, providing a broad spectrum of retinal pathologies for model training and evaluation.

The Papila dataset focuses on the optic nerve head region and contains 420 stereo fundus image pairs (corresponding to 210 subjects). It includes both healthy controls and glaucoma patients, along with corresponding segmentation masks of the optic disc and cup regions.

The Messidor database consists of 1,200 fundus images, originally designed for benchmarking diabetic retinopathy detection algorithms. While the dataset itself lacks fine-grained lesion annotations, the Maples-DR extension [30] provides detailed expert-annotated masks for pathological features including blood vessels, drusen, exudates, hemorrhages, microaneurysms, and cotton wool spots.

To further evaluate the generalizability and robustness of the proposed model, we incorporated the JSIEC dataset as an independent test set. JSIEC contains over 209,000 fundus images representing a wide range of ocular conditions. However, only a limited subset is publicly available via Kaggle [34]. For this study, we selected images labeled as healthy, diabetic retinopathy (DR2), fibrosis, pathological myopia, optic atrophy, and laser spots to cover a representative spectrum of both common and rare retinal pathologies. Using JSIEC solely for testing ensures that the model’s evaluation reflects realistic domain shifts and unseen disease types.

The distribution of fundus images across datasets and diagnostic categories after preprocessing is presented in table 1. This compilation integrates both in-house and public datasets, thereby enabling a comprehensive assessment of classification, localization, and anomaly detection performance.

Table 1: The table shows the distribution of fundus images given the underlying classification and the corresponding dataset after preprocessing.

Classification	AEyeDB	FIVES	Mendeley	Papila	Messidor	JSIEC
Normal	204	178	1024	333	480	38
DR	-	144	1509	-	119	49
Glaucoma	-	99	1349	87	-	-
AMD	-	171	-	-	-	-
Macular Scar (MS)	-	-	444	-	-	-
Retinitis Pigmentosa (RP)	-	-	139	-	-	-
Disc Edema (DE)	-	-	127	-	-	-
Fibrosis	-	-	-	-	-	10
Pathological Myopia	-	-	-	-	-	54

Optic Atrophy	-	-	-	-	-	12
Laser Spots	-	-	-	-	-	20

2.2.2. Fundus Image Preprocessing

The quality of fundus images can vary substantially due to environmental conditions, camera calibration, and subject-specific factors such as eye movement or pupil dilation. These variations often manifest as artifacts including poor illumination, blurring, color distortion, and low contrast, which can impair model performance and reliability. To ensure data consistency, we excluded all low-quality images from the FIVES dataset that were annotated as suboptimal according to the provided quality assessment criteria.

Fundus images also differ with respect to camera angle, field of view, and focal plane, all of which influence the visibility of key intraocular structures such as the optic nerve head, macula, and vascular arcades. Such variations can introduce domain shifts across datasets and affect diagnostic tasks, particularly in diseases like glaucoma, where the optic disc morphology is essential for accurate detection [32].

All images were normalized, center-cropped and resized to 224×224 pixels, aligning with the input resolution of the ViT-B/16 model. To systematically evaluate preprocessing and augmentation strategies for domain generalization, we implemented a stepwise procedure:

1. **Baseline (No Augmentation):** Images were normalized and directly fed into the network without additional transformations to establish a performance reference.
2. **Geometric Augmentation:** We applied random horizontal flipping, full rotations (random degree between 0° and 360°) and random translations up to 10% of image width and height to simulate variations in camera positioning and patient alignment.
3. **Color and Blur Augmentation:** Gaussian blur and color jittering were incorporated to simulate lighting inconsistencies and chromatic aberrations, improving generalization to real-world imaging conditions.
4. **Contrast and Contour Enhancement:** Finally, we evaluated the integration of histogram equalization [37] and Laplacian enhancement [38] within the strongest augmentation setup to emphasize vascular and structural details. These techniques were compared to assess their effect on model convergence and localization quality.

To ensure robust evaluation and prevent overfitting, we randomly reserved 15% of each dataset for validation and 15% for testing, stratified across all disease classes. Figure 2 illustrates examples of the applied preprocessing and augmentation techniques, including geometric transformations, color perturbations, and enhancement methods. The visual examples highlight how the preprocessing pipeline progressively improves the visibility of retinal structures while preserving anatomical integrity.

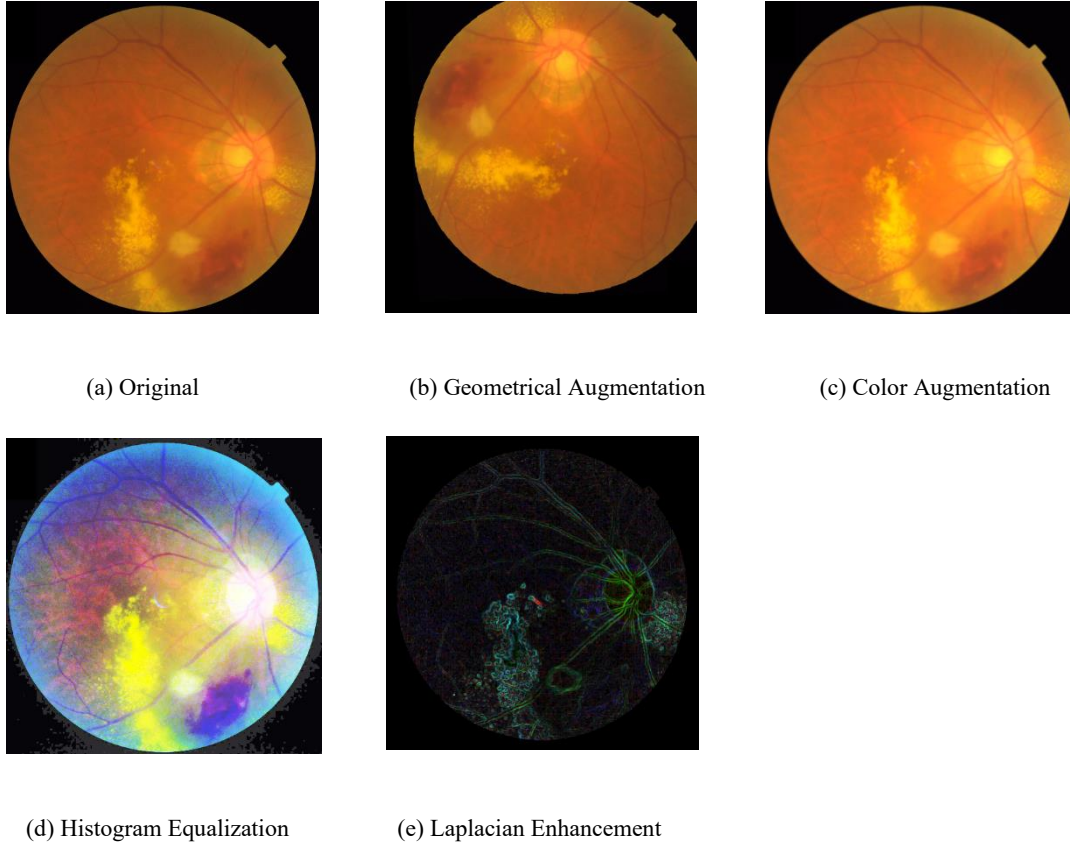


Figure 2: Examples of the preprocessing and augmentation techniques applied to the fundus image datasets. The figure demonstrates successive stages of (a) geometric transformations (random rotation, translation, and flipping), (b) color perturbations (brightness and contrast jittering, Gaussian blur), and enhancement methods (c) histogram equalization and (d) Laplacian filtering). The image was taken from the FIVES dataset.

2.2.3. Training Procedures

To ensure consistency and comparability across datasets, we employed stratified sampling based on both dataset origin and class labels. This approach maintained balanced class distributions within training, validation, and test subsets, preventing potential biases caused by uneven sampling. The training results were stored separately for each dataset, allowing us to quantitatively assess domain shifts and evaluate cross-dataset generalization.

The ViT model was fine-tuned using the Adam optimizer with an initial learning rate of 0.00001 over 30 epochs. A cosine learning rate schedule was applied with five warm-up epochs, enabling gradual adaptation during the early training phase to stabilize convergence. Since the model was initialized with ImageNet-pretrained weights, the model did not require a lot of epochs for fine-tuning.

To ensure full reproducibility, the same random seed was used across all devices, guaranteeing identical data splits and augmentation sequences across experiments. Each preprocessing setup (described in Section 2.2.2) was evaluated independently, ensuring that performance differences arose solely from the applied augmentation or enhancement techniques.

Following the identification of the best-performing preprocessing setup, we trained the constrained attention mechanism as defined in equation 1 using a U-Net based masking network. The pretrained ViT classifier and the U-Net were trained simultaneously which generated an attention mask that

selectively preserves informative regions while suppressing irrelevant background. The model was optimized using the same training configuration as the ViT fine-tuning stage, Adam optimizer with an initial learning rate of 0.00001 and cosine scheduling. This training scheme allows the U-Net to refine its mask output based on the classifier’s gradients, leading to a spatially interpretable representation of regions that contribute most significantly to classification decisions.

In contrast to the pretrained models, the GANomaly network was trained from scratch for 100 epochs, as no pretrained weights were available. The model was optimized using the Adam optimizer with an initial learning rate of 0.00001. The overall loss function consisted of a weighted combination of the reconstruction loss (weighted by 50), the adversarial feature map loss (weighted by 1), the latent space recreation loss (weighted by 1), and the normalized KL regularization (weighted by 0.3). The GANomaly was only trained on the healthy cohort, reducing the number of available samples significantly.

3. Results

3.1 ViT Performance

Overall, the ViT model demonstrated robust classification performance across all tested augmentation strategies, with total accuracies ranging from 0.789 to 0.843 on the test set. Table 2 summarizes key performance metrics for each dataset, providing insight into the model’s generalization capabilities across diverse image sources. Among the evaluated configurations, the geometric augmentation setup achieved the highest overall accuracy (0.843), surpassing both the baseline without augmentation (0.825) and the contrast-enhanced variants. Histogram equalization reached 0.789, while Laplacian enhancement achieved 0.815, both underperforming relative to the non-enhanced baseline. These results suggest that although contrast and edge-enhancement methods can visually emphasize retinal vasculature and optic disc features, they may inadvertently introduce intensity distributions that are non-physiological, thereby slightly degrading model performance. Inclusion of brightness and color augmentations led to marginal changes in raw accuracy (0.841-0.843), but 10-fold cross-validation with a paired t-test confirmed that these differences were not statistically significant ($p > 0.05$), indicating comparable performance across these augmentation schemes. Table 3 presents the multi-class confusion matrix for the full test set. The results indicate that glaucoma and normal images were the most challenging to distinguish, reflecting the subtle structural changes in early-stage glaucoma and the similarity of certain normal anatomical variations to mild pathological features.

Table 2: Final test set performance across various augmentation strategies. Within each cell, results are reported in the following dataset order: AEyeDB, FIVES, Mendeley, Papila, and Messidor. Performance is compared using accuracy, weighted F1-score, and Matthews correlation coefficient (MCC) as evaluation metrics.

	Accuracy	Weighted F1	MCC
No Augmentation	1.000	-	-
	0.842	0.840	0.787
	0.800	0.799	0.732
	0.825	0.786	0.350
	0.944	0.950	0.836

Geometric Augmentation	1.000 0.741 0.839 0.857 0.911	- 0.737 0.839 0.844 0.920	- 0.650 0.784 0.512 0.759
Color Augmentation	1.000 0.730 0.833 0.841 0.956	- 0.722 0.835 0.822 0.961	- 0.634 0.779 0.444 0.865
Histogram Equalization	1.000 0.652 0.769 0.905 0.922	- 0.653 0.770 0.893 0.924	- 0.546 0.692 0.693 0.775
Laplace Enhancement	1.000 0.708 0.798 0.841 0.967	- 0.708 0.801 0.812 0.966	- 0.605 0.736 0.431 0.894

Table 3: The table shows the confusion matrix on the withhold test set for the model trained with geometric augmentation.

	Normal	DR	Glaucoma	AMD	MS	RP	DE
Normal	287	10	34	1	2	0	0
DR	5	245	2	1	6	2	4
Glaucoma	34	10	174	1	10	1	0
AMD	0	2	0	24	0	0	0
MS	3	7	5	1	47	3	1
RP	0	1	0	0	0	20	0
DE	2	2	0	0	0	1	14

3.2 Constrained Attention Mask

Table 4 presents the confusion matrix of the masked ViT model on the test set. Although the integration of the attention-guided U-Net slightly reduced overall performance with an accuracy of 0.832 compared to the baseline ViT model, a 10-fold cross-validation with a paired t-test indicated no statistically significant difference between the two configurations ($p > 0.05$). This suggests that the attention mask functions as intended without adversely affecting classification accuracy.

A detailed experimental comparison of pathological region localization obtained from the attention mask, Grad-CAM, integrated gradients, occlusion analysis, and anomaly detection is provided in Appendix B.

Table 4: The table shows the confusion matrix on the withhold test set for the model trained with geometric augmentation and constrained attention.

	Normal	DR	Glaucoma	AMD	MS	RP	DE
Normal	305	3	23	1	2	0	0
DR	13	229	9	0	9	1	4
Glaucoma	45	8	163	1	12	1	0
AMD	0	3	1	22	0	0	0
MS	3	5	7	1	48	2	1
RP	1	0	0	0	0	20	0
DE	2	2	0	0	0	1	14

3.3 Anomaly Detection

The evaluation of the GANomaly based anomaly detection framework showed that neither the incorporation of KL regularization nor the addition of the attention-mask guided reconstruction loss resulted in measurable performance gains on the test set. Across all configurations, the anomaly detector consistently achieved an AUC of 0.76, indicating comparable ranking performance regardless of regularization strategy. However, their training dynamics differed: introducing stochasticity via the KL term in combination with the adversarial objective produced visibly noisier and less stable loss trajectories, as illustrated in Figure 3a, whereas deterministic configurations converged more smoothly.

Figure 3b presents the distribution of anomaly scores from the KL-regularized model. These scores were subsequently transformed into calibrated abnormality probabilities using GUESS, which applies Bayesian posterior estimation to map raw anomaly magnitudes to interpretable probability values. This calibration step enables the model to output continuous estimates reflecting the likelihood of pathology, thereby avoiding reliance on discrete threshold selection.

Given the objective of producing well-ordered probabilistic predictions we assessed performance primarily using the AUC, which measures ranking quality rather than threshold-specific accuracy. This metric is particularly appropriate for anomaly detection systems and supports clinical use cases where decision thresholds may vary depending on risk tolerance or screening context.

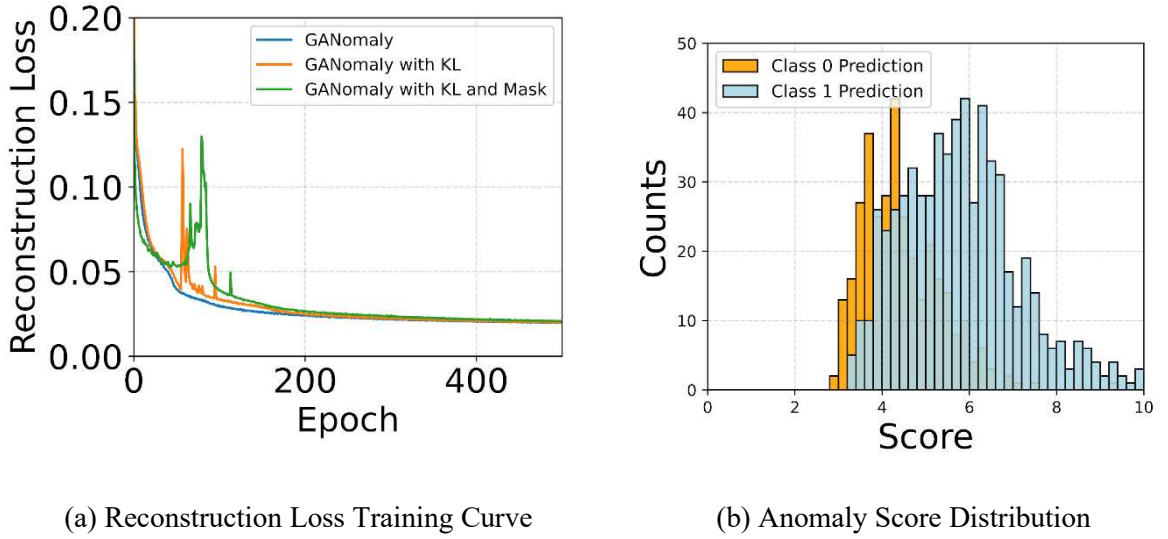


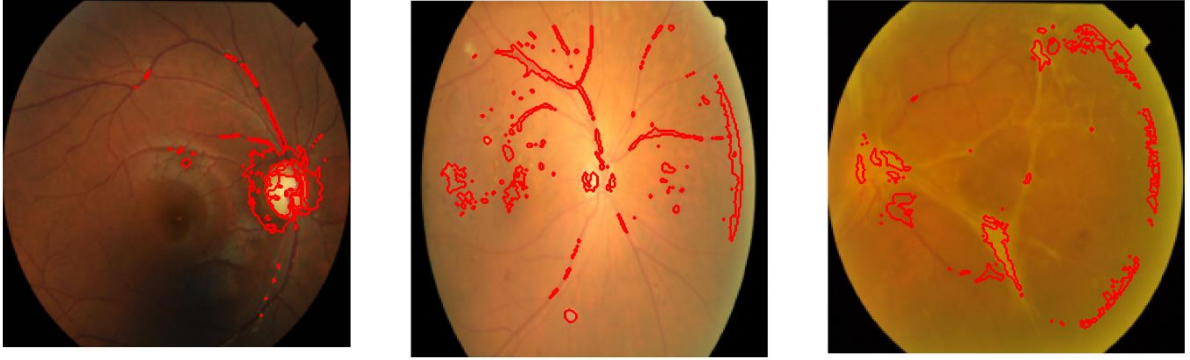
Figure 3: The reconstruction loss training curves for the three setups are shown in figure (a) while figure (b) shows the distribution of the uncalibrated anomaly score on the test set.

A comparative evaluation on the unseen JSIEC dataset is summarized in table 5, contrasting the performance of the baseline ViT classifier with that of the GANomaly anomaly detector. For the ViT model, class-level predictions are reported directly, whereas for the GANomaly model, we present the mean calibrated abnormality probabilities obtained through GUESS for each disease category. This allows for a direct comparison between deterministic class assignments and probabilistically derived anomaly estimates.

Figure 4 visualizes the regions contributing most strongly to the anomaly signal by displaying the 90th-percentile absolute reconstruction errors for several pathological patterns. These maps highlight the correspondence between elevated reconstruction error and clinically relevant structures, offering qualitative insight into the detector’s failure modes and emphasizing its potential utility as a weakly supervised localization tool.

Table 5: The table provides the corresponding fractions of classification for the ViT base model and the mean probabilities of abnormalities for the GANomaly model calibrated with GUESS.

	Base Model	Mean Probability of an Anomaly for GANomaly
Normal	Normal: 1.00	0.71±0.09
DR	DR: 1.00	0.76±0.12
Fibrosis	AMD: 0.20 DR: 0.70 MS: 0.10	0.81±0.11
Pathological Myopia	AMD: 0.35 DR: 0.26 Glaucoma: 0.09 MS: 0.30	0.94±0.07
Optic Atrophy	Normal: 0.17 DR: 0.17 Glaucoma: 0.66	0.76±0.05
Laser Spots	Normal: 0.05 DE: 0.05 DR: 0.45 Glaucoma: 0.05 RP: 0.4	0.87±0.07



(a) Healthy Image

(b) DR Image

(c) Fibrosis Image

Figure 4: Visualization of the 90th-percentile absolute reconstruction error maps for representative pathological cases. These maps highlight the image regions contributing most strongly to the anomaly signal, illustrating the correspondence between elevated reconstruction error and clinically relevant structures.

4. Discussion

The performance of the ViT across multiple augmentation strategies highlights both the potential and the current challenges of fundus-based disease classification. Although the overall accuracy remains high across all configurations, the observed variability in weighted F1 and MCC underscores the difficulty of achieving balanced performance across all disease categories, an expected outcome in inherently imbalanced medical datasets .

Early-stage disease detection remains particularly challenging, as reflected in the confusion matrix. While the ViT model reliably distinguishes healthy from pathological images, it exhibits class-specific variability. DR and AMD classes demonstrate strong sensitivity and consistent classification performance, whereas glaucoma shows the highest rate of misclassification, often being predicted as normal. This observation suggests that early glaucomatous changes, which primarily manifest as subtle alterations in the optic nerve head and peripapillary regions, remain difficult to capture even for high-capacity transformer-based models.

Interestingly, no single augmentation strategy yielded a statistically significant improvement across all datasets. Both geometric and color augmentation produced comparable results, suggesting that the model effectively leverages both spatial and chromatic cues. However, histogram equalization achieved the highest accuracy within the Papila dataset, composed exclusively of normal and glaucoma images, indicating that contrast normalization may enhance feature discrimination when subtle textural and structural variations dominate. In contrast, Laplacian enhancement and heavy color perturbations tended to reduce performance, likely due to oversharpening and the artificial amplification of irrelevant image gradients.

When evaluating the ViT model with geometric augmentation on the Papila dataset, the model achieved an AUC of 0.91. This represents a substantial improvement compared to the results reported in the original Papila study [33], which achieved an AUC of approximately 0.87 using a conventional

convolutional network ensemble. The increase can be attributed to the use of a state-of-the-art transformer-based architecture, the integration of multiple datasets during training which enriches the learned feature distribution, and the application of effective data augmentation strategies that improve generalization across unseen samples.

Differences in dataset-level performance further highlight the impact of domain shift, as illustrated by the training curves presented in Appendix A. Datasets characterized by consistent acquisition protocols and manual quality control, such as FIVES and AEyeDB, achieved the highest performance, indicating that standardized imaging conditions and reduced intra-dataset variability substantially enhance model generalization. In contrast, the particularly strong results on AEyeDB may also reflect overfitting to dataset-specific image characteristics, emphasizing the need for future work on cross-dataset normalization and domain adaptation techniques.

Although the constrained attention mechanism did not improve classification accuracy, its integration represents a meaningful step toward model-intrinsic interpretability. Unlike post-hoc methods such as Grad-CAM or occlusion, which rely on surrogate explanations that may be biased by baseline image choices [39], our U-Net derived spatial masks are optimized jointly with the classifier. Despite the minor performance decrease, this approach offers a more principled way of enforcing spatial consistency and may benefit clinical interpretability in future extensions.

The GANomaly-based anomaly detector achieved a consistent AUC of 0.76, independent of whether KL regularization or mask-based reconstruction loss was applied. Unlike supervised classifiers, anomaly detectors are mathematically grounded to generalize to unseen pathology distributions. The calibrated probabilities obtained through GUESS provide interpretable outputs and avoid threshold biases. Moreover, the reconstruction-error maps highlight clinically meaningful structures, offering an inherently explainable signal that points to suspicious regions without requiring class labels.

Unlike supervised classifiers, anomaly detectors are mathematically grounded to generalize to unseen pathology distributions. The calibrated probabilities obtained through GUESS provide interpretable outputs and avoid threshold biases. Moreover, the reconstruction-error maps highlight clinically meaningful structures, offering an inherently explainable signal that points to suspicious regions without requiring class labels.

Future work should prioritize increasing the number of healthy training images, as the current imbalance biases the probabilistic calibration toward higher abnormality likelihoods. Additionally, domain adaptation and dataset harmonization will be essential to mitigate dataset-specific biases.

5. Conclusion

This work demonstrates the strong potential of transformer-based architectures for retinal disease classification while emphasizing enduring limitations related to subtle pathology detection and cross-domain robustness. The ViT achieved high performance across multiple datasets and augmentation strategies, with particularly reliable detection of DR and AMD. Glaucoma remained the most difficult class, reflecting the subtle morphological changes associated with early disease. No augmentation strategy yielded a universal improvement. Geometric and color augmentations were consistently beneficial, whereas Laplacian enhancement reduced discriminability probably due to oversharpening artifacts. Histogram equalization was advantageous only in datasets dominated by nuanced structural

cues, such as Papila. Our cross-dataset evaluation revealed the substantial influence of acquisition quality and dataset consistency on model generalization.

The GANomaly-based anomaly detector, although less accurate than the ViT due to limited training diversity and lower model capacity, offered two distinct advantages: intrinsic explainability through reconstruction-based error maps and stable generalization to unseen datasets. Its GUESS-calibrated probabilistic outputs avoid threshold tuning and provide a mathematically grounded alternative to classifier confidence scores. This makes anomaly detection a valuable complementary strategy, especially in settings where disease categories are open-ended or incompletely represented during training.

Future work will focus on expanding the pool of healthy fundus images to improve calibration stability, reducing domain shift through harmonization and self-supervised pretraining, and exploring hybrid architectures that combine discriminative transformer models with generative anomaly detection. Together, these directions aim to advance trustworthy, explainable, and generalizable AI tools for ophthalmic screening.

Ethics Statement

All research complied with German legal regulations. The study was approved by the Ethics Committee of the Heinrich Heine University on November 8, 2023 (ID: 2023-2622). Informed consent was obtained from all participants.

Acknowledgements

This work is partially funded by hessian.AI in the Connectom project VirtualDoc. We also wish to thank the students of the Heinrich Heine University Düsseldorf who participated in the study. The MESSIDOR dataset was kindly provided by the Messidor program partners.

Author Contributions

D. H. served as the principal investigator, overseeing the study’s design and implementation. B. F. and T. P. provided valuable insights during the analysis phase. J. B. R. led the study’s execution, analysed the data, and conceptualized the study. A. C., J. S., and N. E. contributed to the ethical proposal, study implementation, and medical interpretation.

Data Availability Statement

As outlined in the study description (available: <https://drks.de/search/de/trial/DRKS00033094>), we are committed to adhering to the data security policies of both the European Union (EU) and Germany. Accordingly, any researcher interested in accessing the data may do so by submitting valid research comply with the data security regulations of both the EU and Germany. This process ensures the confidentiality and integrity of the data while facilitating responsible and ethical research practices

References

- [1] T.-Y. Wong, U. Chakravarthy, R. Klein, P. Mitchell, G. Zlateva, R. Buggage, K. Fahrback, C. Probst, and I. Sledge, “The natural history and prognosis of neovascular age-related macular

degeneration. a systematic review of the literature and meta-analysis,” *Ophthalmology*, vol. 115, pp. 116–26, 01 2008.

[2] M. Fleckenstein, T. Keenan, R. Guymer, U. Chakravarthy, S. Schmitz-Valckenberg, C. Klaver, W. Wong, and E. Chew, “Age-related macular degeneration,” *Nature Reviews Disease Primers*, vol. 7, p. 31, 05 2021.

[3] D. Antonetti, P. Silva, and A. Stitt, “Current understanding of the molecular and cellular pathology of diabetic retinopathy,” *Nature Reviews Endocrinology*, vol. 17, pp. 1–12, 01 2021.

[4] R. Casson, G. Chidlow, J. Wood, J. Crowston, and I. Goldberg, “Definition of glaucoma: Clinical and experimental concepts,” *Clinical & experimental ophthalmology*, vol. 40, pp. 341–9, 02 2012.

[5] K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, and J. Ye, “Fives: A fundus image dataset for ai-based vessel segmentation,” 2022.

[6] L.-P. Cen, J. Ji, J.-W. Lin, S.-T. Ju, H.-J. Lin, T.-P. Li, Y. Wang, J.-F. Yang, Y.-F. Liu, S. Tan, L. Tan, D. Li, Y. Wang, D. Zheng, Y. Xiong, H. Wu, J. Jiang, Z. Wu, D. Huang, T. Shi, B. Chen, J. Yang, X. Zhang, L. Luo, C. Huang, G. Zhang, Y. Huang, T. K. Ng, H. Chen, W. Chen, C. P. Pang, and M. Zhang, “Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks,” *Nature Communications*, vol. 12, Aug. 2021.

[7] J. Guo, Z. Pang, F. Yang, J. Shen, and J. Zhang, “Study on the method of fundus image generation based on improved gan,” *Mathematical Problems in Engineering*, vol. 2020, pp. 1–13, 07 2020.

[8] S. Go, Y. Ji, S. J. Park, and S. Lee, “Generation of structurally realistic retinal fundus images with diffusion models,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, p. 2335–2344, IEEE, June 2024.

[9] Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018 Mar;2(3):158-164. doi: 10.1038/s41551-018-0195-0. Epub 2018 Feb 19. PMID: 31015713.

[10] Mitani, A., Huang, A., Venugopalan, S. *et al.* Detection of anaemia from retinal fundus images via deep learning. *Nat Biomed Eng* 4, 18–27 (2020). <https://doi.org/10.1038/s41551-019-0487-z>

[11] Zhu Z, Shi D, Guankai P, Tan Z, Shang X, Hu W, Liao H, Zhang X, Huang Y, Yu H, Meng W, Wang W, Ge Z, Yang X, He M. Retinal age gap as a predictive biomarker for mortality risk. *Br J Ophthalmol*. 2023 Apr;107(4):547-554. doi: 10.1136/bjophthalmol-2021-319807. Epub 2022 Jan 18. PMID: 35042683.

[12] Yurdakul, M., Uyar, K., and Tasdemir, S., “MaxGlaViT: A novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images”, Art. no. arXiv:2502.17154, 2025. doi:10.48550/arXiv.2502.17154.

[13] Li Z, Han Y, Yang X. Multi-Fundus Diseases Classification Using Retinal Optical Coherence Tomography Images with Swin Transformer V2. *Journal of Imaging*. 2023; 9(10):203. <https://doi.org/10.3390/jimaging9100203>

[14] Hu W, Li K, Gagnon J, Wang Y, Raney T, Chen J, Chen Y, Okunuki Y, Chen W, Zhang B. FundusNet: A Deep-Learning Approach for Fast Diagnosis of Neurodegenerative and Eye Diseases

- [15] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017.
- [16] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” 2017.
- [17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, June 2016.
- [18] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” arXiv e-prints, p. arXiv:1703.01365, Mar. 2017.
- [19] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, “GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training,” arXiv e-prints, p. arXiv:1805.06725, May 2018.
- [20] Johanna Schwarz, Dominik Heider, GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making, *Bioinformatics*, Volume 35, Issue 14, July 2019, Pages 2458–2465, <https://doi.org/10.1093/bioinformatics/bty984>
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” arXiv e-prints, p. arXiv:1706.03762, June 2017.
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv e-prints, p. arXiv:2010.11929, Oct. 2020.
- [23] Q. Bi, X. Sun, S. Yu, K. Ma, C. Bian, M. Ning, N. He, Y. Huang, Y. Li, H. Liu, and Y. Zheng, “Mil-vit: A multiple instance vision transformer for fundus image classification,” *Journal of Visual Communication and Image Representation*, vol. 97, p. 103956, 2023.
- [24] X. Ouyang *et al.*, “Learning Hierarchical Attention for Weakly-Supervised Chest X-Ray Abnormality Localization and Diagnosis,” in *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2698–2710, Oct. 2021, doi: 10.1109/TMI.2020.3042773.
- [25] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang, S. Xu, and H. Chen, “An improved transformer network for skin cancer classification,” *Computers in Biology and Medicine*, vol. 149, p. 105939, 2022.
- [26] Q. Bi, X. Sun, S. Yu, K. Ma, C. Bian, M. Ning, N. He, Y. Huang, Y. Li, H. Liu, and Y. Zheng, “Mil-vit: A multiple instance vision transformer for fundus image classification,” *Journal of Visual Communication and Image Representation*, vol. 97, p. 103956, 2023.
- [27] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., “Learning Deep Features for Discriminative Localization”, arXiv:1512.04150, 2015. doi:10.48550/arXiv.1512.04150.
- [28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, p. 336–359, Oct. 2019.

- [29] Jan Benedikt Ruhland, Iraj Masoudian, Dominik Heider, Enhancing deep neural network training through learnable adaptive normalization, *Knowledge-Based Systems*, Volume 326, 2025, 113968, ISSN 0950-7051, <https://doi.org/10.1016/j.knosys.2025.113968>.
- [30] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1312.6114>
- [31] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695. <https://arxiv.org/abs/2112.10752>
- [32] Shayla Sharmin, Mohammad Riadur Rashid, Tania Khatun, Md Zahid Hasan, Mohammad Shorif Uddin, Marzia, A dataset of color fundus images for the detection and classification of eye diseases, *Data in Brief*, Volume 57, 2024, 110979, ISSN 2352-3409, <https://doi.org/10.1016/j.dib.2024.110979>.
- [33] Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R. et al. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Sci Data* **9**, 291 (2022). <https://doi.org/10.1038/s41597-022-01388-1>
- [34] E. Decenci re, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, “Feedback on a publicly distributed image database: The messidor database,” *Image Analysis and Stereology*, vol. 0, 07 2014.
- [35] L. Chundan, “1000 Fundus Images (39 categories) – Joint Shantou International Eye Centre (JSIEC),” *Kaggle Dataset*, 2020. [Online]. Available: <https://www.kaggle.com/datasets/linchundan/>
- [36] G. Lepetit-Aimon, C. Ployout, M. Boucher, R. Duval, M. Brent, and F. Cheriet, “Maples-dr: Messidor anatomical and pathological labels for explainable screening of diabetic retinopathy,” *Scientific Data*, vol. 11, 08 2024.
- [37] Manpreet Kaur, Jasdeep Kaur, Jappreet Kaur, “Survey of Contrast Enhancement Techniques based on Histogram Equalization,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 2, no. 7, 2011.
- [38] X. Ren and S. Lai, “Medical Image Enhancement Based on Laplace Transform, Sobel Operator and Histogram Equalization,” *Academic Journal of Computing & Information Science*, vol. 5, no. 6, pp. 48-54, 2022.
- [39] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

Appendix A. Network Design and Ablation Details

A.1. Network Design

Table B.6: Architectural summary of the proposed U-Net model used for constrained attention mask generation. The network integrates residual blocks with attention at the middle section, featuring symmetric encoder-decoder paths with skip connections. Downsampling and upsampling operations control spatial resolution, while a final sigmoid activation constrains the output mask to the range $[0,1]$.

Stage	Layer / Block	Input Channels	Output Channels	Key Features / Notes
Input	AdaptiveNorm + Conv2d	3	32	Input normalization, 3×3 conv, padding=1
Down 1	Downsample	32	32	Residual connections
Down 2	Downsample	32	64	Residual block
Down 3	Downsample	64	128	Residual block
Down 4	Identity	128	128	No downsampling
Middle	ResBlockWithAtt	128	128	Attention always applied, residual connections
Up 4	Upsample	256 (skip+up)	128	Concatenates skip connection from Down 4
Up 3	Upsample	256	64	Concatenates skip connection from Down 3
Up 2	Upsample	128	32	Concatenates skip connection from Down 2
Up 1 (last)	Identity	64	32	Concatenates skip connection from Down 1
Output	Conv2d + Sigmoid	32	1	Produces single-channel mask, pixel values in $[0,1]$

A.2. Ablation

Figure B.5 illustrates the learning curves for the different augmentation strategies. In figure B.6 the accuracies on the validation set for each individual dataset is listed.

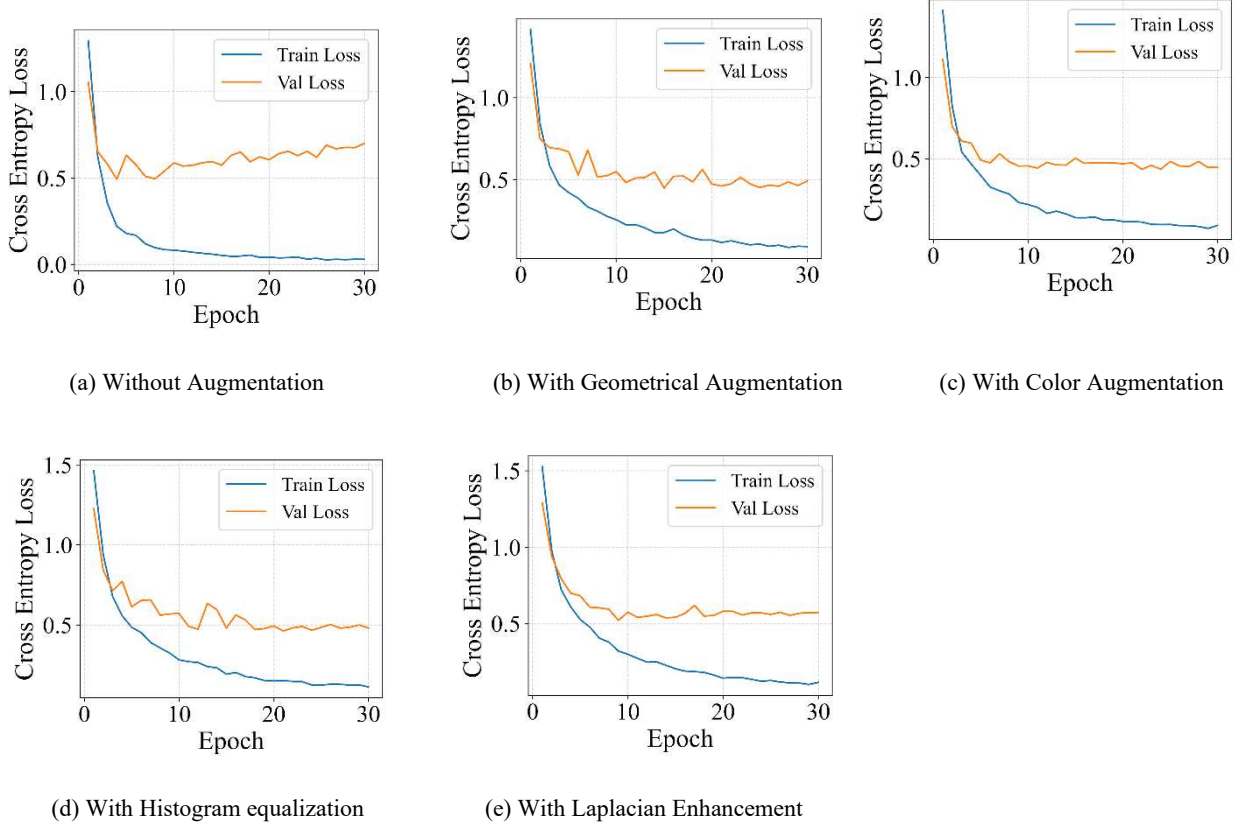


Figure B.5: Learning Curves for different augmentation strategies.

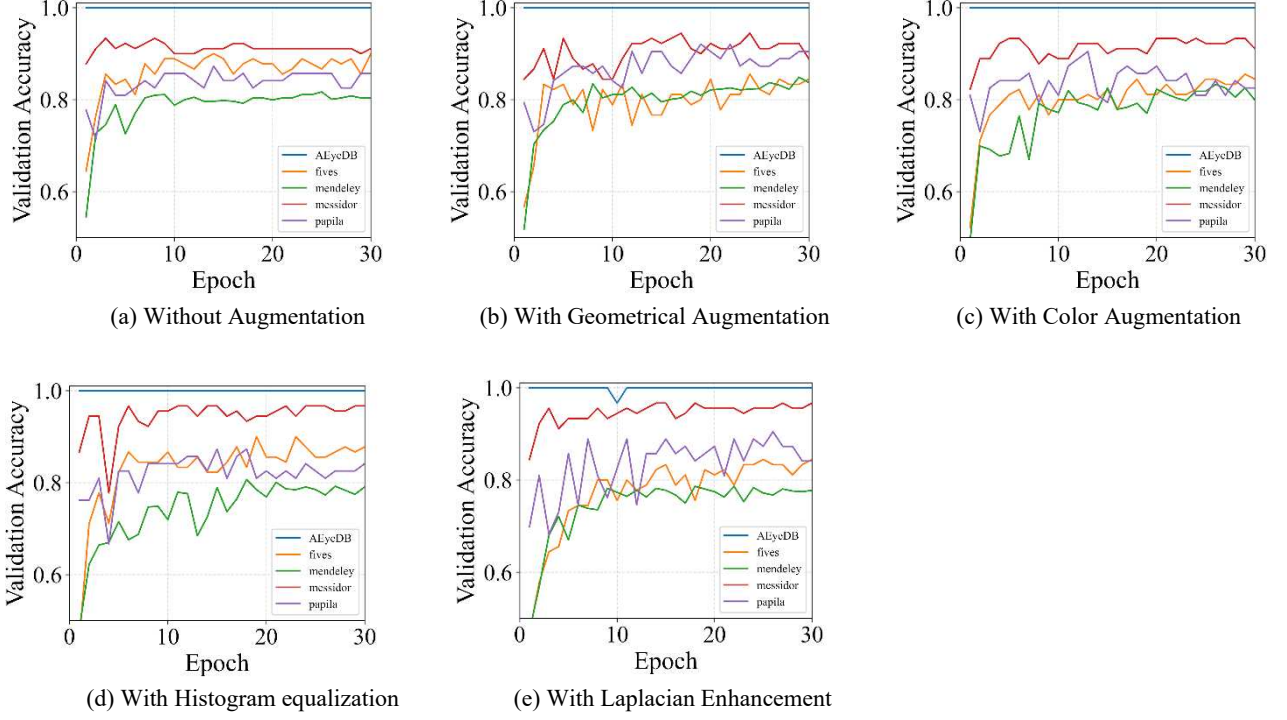


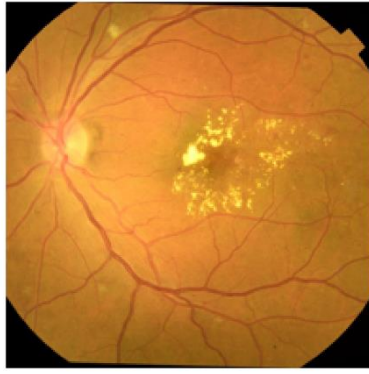
Figure B.6: The figure illustrates the validation accuracy per dataset to observe any domain shifts.

Appendix B. Model Explainability

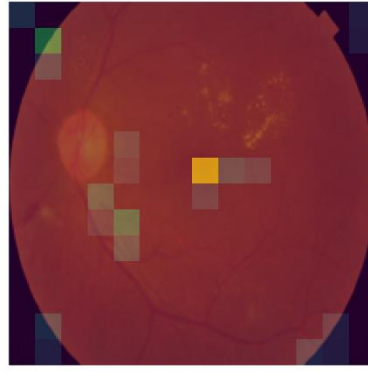
Figure B.7 presents a comparison of several explainable AI techniques, highlighting their respective limitations when applied to fundus image interpretation. The first method evaluated was patch-based occlusion, in which image regions are systematically masked and the resulting changes in model prediction are measured. We report the 90th-percentile impact scores across the two patch sizes 16×16 (matching the ViT patch resolution) and 8×8 (providing finer spatial granularity). However, the resulting saliency patterns exhibit highly erratic and spatially inconsistent behavior, limiting the clinical interpretability and reliability of this approach.

We further assessed integrated gradients and Grad-CAM, which derive localization from gradient-based changes. While CAM is well suited for CNNs with designated feature aggregation layers, its application to ViTs is less straightforward because no single layer naturally encodes spatial importance, leading to unstable or layer-dependent results. SHAP-like methods present an additional challenge: they require a predefined baseline, and simple choices such as black or white images may introduce unintended biases or unrealistic reference states.

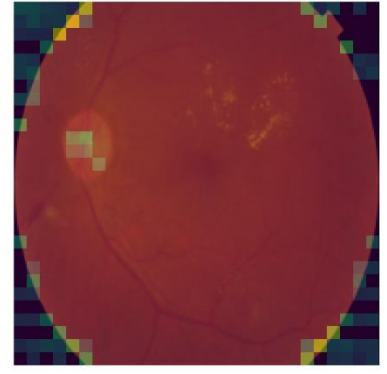
For comparison, we also include the constrained attention mask proposed in this work. Unlike post-hoc explainable AI methods, it provides a stable and model-intrinsic representation of spatial relevance that is learned jointly during training.



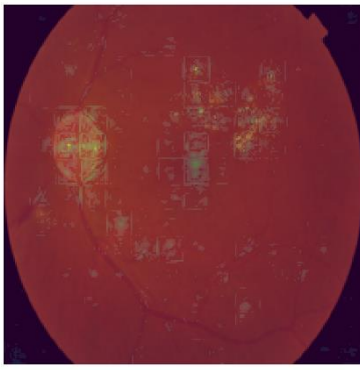
(a) Original



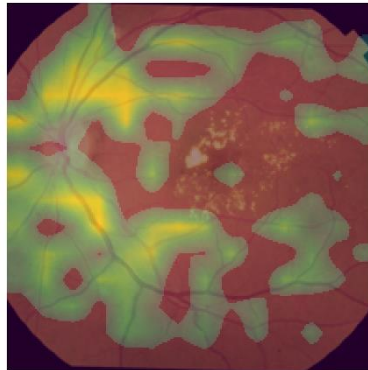
(b) 8×8 Occlusion



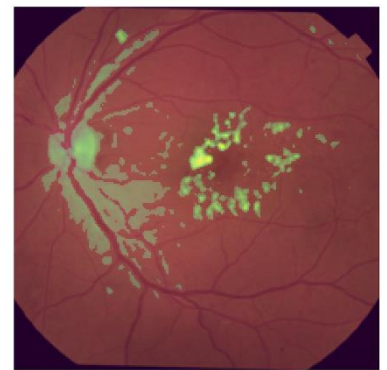
(c) 16×16 Occlusion



(d) Integrated Gradients



(e) Grad-CAM



(f) Attention Mask

Figure B.7: Comparison of explainable AI methods including occlusion analysis (8×8 and 16×16 patches), gradient-based CAM, integrated gradient contribution, and the proposed constrained attention mask. The visualization highlights the instability and methodological limitations of traditional post-hoc approaches, in contrast to the consistency of the integrated attention-based method. The image was taken from the Mendeley dataset and contains DR pathologies.

4.4. Development of the Virtual Doctor Avatar and Chatbot

The final project presented in this thesis focused on the development of the Virtual Doctor Avatar, which aimed to provide a human-like, interactive interface for patients engaging with the Virtual Doctor system. The overarching objective was to design a conversational agent capable of conducting an automated yet targeted anamnesis in a natural and user-friendly manner. To achieve this, the avatar needed to combine advanced multimodal processing components that would allow it not only to understand and respond to spoken language but also to embody a visually realistic representation of a physician-like figure.

To realize this vision, multiple models were trained and integrated into the avatar framework. These models were responsible for different modalities essential to the functioning of the avatar. Specifically, a text-to-speech component was implemented to generate natural and context-appropriate spoken responses, thereby enhancing the sense of realism and engagement in the human-machine interaction. In addition, a conditional diffusion model was explored, enabling the dynamic generation of facial animations for the avatar that synchronized with the spoken output. Together, these components created a cohesive system in which the avatar not only provided verbal feedback but also conveyed human-like presence through synchronized audiovisual cues.

The targeted anamnesis conducted by the avatar was supported by a fine-tuned large language model trained on medical data. This model was capable of asking contextually relevant follow-up questions, thereby extending beyond static questionnaires and approximating the dialogic nature of real clinical anamnesis. Importantly, the system was designed to avoid providing explicit diagnostic statements to patients; instead, it generated a structured summary of the anamnesis, which was stored securely in the database and made available exclusively to the corresponding physicians. This ensured that the avatar supported the diagnostic workflow while maintaining compliance with ethical and legal requirements for medical decision-making.

4.4.1. Contributions

I was primarily responsible for the conceptualization, design, and implementation of the avatar component, including the integration of the text-to-speech and image generation models into a unified pipeline. While the underlying chatbot architecture was developed by a colleague, I provided close supervision and guidance throughout its creation, ensuring methodological rigor and compatibility with the avatar framework. Furthermore, I conducted systematic testing and refinement of the integrated system to optimize usability and reliability in real-world conditions.

The Principal Investigator of the project was Prof. Dr. Dominik Heider, who confirms that the description of my contributions is accurate.

Journal:	Computers in Biology and Medicine
Contribution to the publication:	70%
Type of authorship:	First Author
Status of publication:	Submitted
Preprint:	https://arxiv.org/abs/2511.18632

Münster, 25.11.2025

Place, Date



Prof. Dr. Dominik Heider

The Locally Deployable Virtual Doctor: LLM Based Human Interface for Automated Anamnesis and Database Conversion

Jan Benedikt Ruhland¹, Doguhan Bahcivan¹, Jan-Peter Sowa², Ali Canbay², Dominik Heider^{3,4,*}

1: Heinrich Heine University Düsseldorf, Faculty of Mathematics and Natural Sciences, Düsseldorf, Germany

2: University Hospital Knappschaftskrankenhaus Bochum, Department of Internal Medicine, Bochum, Germany

3: Philipps University of Marburg, Department of Mathematics and Computer Science, Marburg, Germany

4: University of Münster, Institute of Medical Informatics, Münster, Germany

*corresponding author: dominik.heider@uni-muenster.de

Abstract

Recent advances in large language models made it possible to achieve high conversational performance with substantially reduced computational demands, enabling practical on-site deployment in clinical environments. Such progress allows for local integration of AI systems that uphold strict data protection and patient privacy requirements, yet their secure implementation in medicine necessitates careful consideration of ethical, regulatory, and technical constraints.

In this study, we introduce MedChat, a locally deployable virtual physician framework that integrates an LLM-based medical chatbot with a diffusion-driven avatar for automated and structured anamnesis. The chatbot was fine-tuned using a hybrid corpus of real and synthetically generated medical dialogues, while model efficiency was optimized via Low-Rank Adaptation. A secure and isolated database interface was implemented to ensure complete separation between patient data and the model's inference process. The avatar component was realized through a conditional diffusion model operating in latent space, trained on researcher video datasets and synchronized with mel-frequency audio features for realistic speech and facial animation.

Unlike existing cloud-based systems, this work demonstrates the feasibility of a fully offline, locally deployable LLM-diffusion framework for clinical anamnesis. The autoencoder and diffusion networks exhibited smooth convergence (final $L1 = 0.037$, $MSE = 0.016$), and MedChat achieved stable fine-tuning with strong generalization to unseen data. The proposed system thus provides a privacy-preserving, resource-efficient foundation for AI-assisted clinical anamnesis, also in low-cost settings.

Keywords: Machine Learning, AI in medicine, LLM, Diffusion Model, Autoencoder

1. Introduction

Artificial intelligence (AI) has reached a pivotal moment in its development, leading to transformative changes across nearly all sectors of human society [1-4]. The rapid evolution of AI technologies presents both significant opportunities and substantial challenges, particularly with respect to adaptation, regulation, and ethical implementation. Central to this advancement is the emergence and continuous improvement of Large Language Models (LLMs), which are not only driving innovation in natural language processing and human-computer interaction but also in other fields [2, 5].

One prominent example is ChatGPT [6], an easily accessible and highly versatile LLM that has demonstrated utility across a wide range of applications. These include grammar correction, content refinement, application development, and even educational support. Despite their flexibility and impressive capabilities, such models possess significant risks. They may produce incorrect or misleading outputs, especially when their responses are accepted without critical scrutiny [7]. Nevertheless, the societal potential of LLMs remains vast, particularly when appropriately integrated into domain-specific workflows.

In the field of medicine, AI occupies a uniquely sensitive and impactful position. The application of machine learning to medical data requires rigorous considerations due to the inherently private, complex, and ethically sensitive nature of healthcare information [8, 9]. Ensuring data privacy, maintaining transparency, and adhering to ethical research practices are essential prerequisites that complicate the implementation of AI in clinical settings. However, despite these challenges, AI has already demonstrated its capacity to augment human expertise. For instance, convolutional neural networks have been reported to show superior performance in certain diagnostic tasks, particularly in medical imaging and pattern recognition [10, 11].

One area where AI has the potential to further improve healthcare delivery is in anamnesis, the medical history-taking process. Effective anamnesis is complicated by the multivariate nature of symptoms, disease-specific manifestations, and the potential for asymptomatic progression, as seen in conditions such as COVID-19 [12]. Traditional approaches often rely on standardized questionnaires, which may overlook critical individual nuances. In contrast, LLMs offer the flexibility to adapt questions dynamically based on patient responses, allowing for a more personalized and comprehensive data collection process.

By integrating LLMs into anamnesis workflows, clinicians could receive more structured, relevant, and individualized information, potentially leading to earlier and more accurate diagnoses. Much like the role of CNNs in medical image tasks, LLMs could serve as decision-support tools that assist healthcare professionals, reduce the risk of diagnostic errors, and contribute to the early identification of severe diseases. Thus, with appropriate safeguards, AI holds promise as a transformative aid in the practice of modern medicine.

As demonstrated in a recent study by Tu et al., LLMs have begun to outperform medical professionals in real-world clinical settings [13]. Their model, the Articulate Medical Intelligence Explorer (AMIE), was evaluated in a controlled study where trained actors simulated patient interactions, allowing for direct comparison between AMIE and practicing clinicians during the anamnesis process. The objective of the study was to assess the viability of deploying a domain-specific LLM for clinical use, particularly in patient history-taking. The results revealed that AMIE, even when used by non-experts, outperformed licensed medical professionals across multiple diagnostic dimensions. This underscores the

transformative potential of LLMs in supporting or even enhancing the clinical decision-making process, especially in initial diagnostic stages such as anamnesis.

Our study aimed to develop a locally deployable LLM, named MedChat, optimized for use in real-world clinical environments with constrained hardware capabilities. A key objective was to ensure secure integration with patient data systems. To that end, MedChat was designed to access a patient's medical history only through a pre-processing interface: the relevant data is extracted and securely fed to the model prior to the interaction, without granting the LLM direct access to the underlying database. This approach balances the benefits of personalized context with stringent privacy and security standards. In addition, safeguards were implemented to mitigate risks associated with misuse or unauthorized manipulation of the chatbot, addressing a common concern in deploying AI systems in sensitive domains such as healthcare.

Beyond the development of the LLM itself, we also designed a human-machine interface to enhance the interaction quality and user engagement. To achieve this, video recordings of two researchers were used to train a diffusion model operating on latent space transformations, capable of generating synchronized speech and facial animations based on textual input. This enabled the creation of a virtual physician avatar, effectively simulating a human-like presence without requiring real-time human supervision. The resulting system offers a fully automated, responsive, and lifelike clinical interaction environment, potentially serving as a scalable solution for digital anamnesis.

The main contributions of this work are summarized as follows:

1. We present MedChat, a locally deployable large language model tailored for clinical anamnesis, capable of operating fully offline on consumer-grade hardware.
2. We integrate a diffusion-based virtual physician avatar for naturalistic, multimodal interaction combining speech and facial animation.
3. We employ parameter-efficient fine-tuning (LoRA) to achieve medical adaptation while maintaining the general-purpose capabilities of the base model.
4. We propose a secure database interface that enables structured, privacy-preserving integration of patient information without granting direct database access to the language model.
5. We demonstrate the complete end-to-end training and evaluation of all system components, laying the groundwork for future clinical validation.

Research in Context

The integration of LLMs into healthcare has advanced rapidly, encompassing diverse applications ranging from diagnostic assistance and clinical documentation support to systems that approach partial or near-full autonomy in tasks traditionally performed by physicians, such as patient consultation and treatment planning [13, 14]. Current implementations include clinical decision support systems [14], symptom-based triage agents [15], and automated virtual health advisors [16]. More experimental research has focused on developing fully autonomous “Virtual Doctor” systems capable of interacting directly with patients; however, these remain largely at the conceptual or prototype stage due to persistent ethical, regulatory, and safety challenges [17].

A major limitation in the deployment of LLMs and their embodied counterparts lies in their substantial computational requirements. High-performance models demand extensive memory resources and GPU processing capacity [18], often making cloud-based deployment necessary. While such infrastructure alleviates the need for on-premise hardware, it introduces significant concerns related to privacy, data protection, and compliance with medical data governance frameworks, particularly when sensitive information such as patient records, imaging data, or personal health histories is processed.

Most existing LLM-based healthcare systems rely on purely text-based chatbot interfaces that lack visual embodiment or avatar representation [13-16]. Although these models provide robust functional capabilities, their absence of visual and emotional expressiveness can reduce patient engagement and trust, elements that are fundamental in clinical communication and adherence. To address this limitation, recent research has investigated the integration of avatars and virtual agents capable of visual and nonverbal interaction [19, 20].

Unlike avatars developed through game-engine-based real-time rendering or pre-recorded video playback, AI-driven avatar systems are typically implemented using generative approaches such as Generative Adversarial Networks [21] or diffusion-based models [22]. These architectures enable the synthesis of realistic facial expressions, gestures, and lip-synchronized speech, offering the potential for more natural and immersive interactions. However, achieving real-time or near-real-time responsiveness remains a considerable technical challenge. Full-scale diffusion models exhibit high latency, making them unsuitable for dynamic, interactive clinical environments. Consequently, researchers have explored lightweight alternatives, including optimized GANs and parameter-efficient diffusion models, often referred to as “fast” or “turbo” diffusion variants, that aim to balance visual fidelity with computational efficiency [23, 24].

Despite these advancements, avatar-based or visually embodied LLM systems continue to face substantial computational barriers. These architectures require not only sufficient GPU memory to host the LLM and associated generative modules, often exceeding 12-16 GB for high-resolution outputs, but also high-throughput tensor processing to meet the latency constraints of real-time interaction. Accordingly, the hardware and performance demands of such systems remain a significant obstacle to their large-scale deployment and integration within clinical workflows.

In contrast to prior work, our approach uniquely combines local deployability, multimodal avatar-based communication, and strict data isolation mechanisms. Whereas systems such as AMIE or MedAssist rely on cloud infrastructure with extensive computational demands, MedChat demonstrates that comparable performance can be achieved under resource constraints while maintaining full data sovereignty and GDPR (General Data Protection Regulation) compliance. Table 1 provides a comparison of our model MedChat and other existing systems.

Table 1: Comparison of existing LLM-based medical systems.

System	Deployment	Avatar Interface	Privacy Model	Clinical Readiness
AMIE [13]	Cloud	No	Limited	Experimental
MedAssist [14]	Cloud	No	Partial	Prototype
HealsHealthAI [16]	Cloud	No	Partial	Research
MedChat	Local	Yes	Full Isolation	Preclinical Evaluation Planned

2. Materials and Methods

2.1 Project Overview

The primary objective of this project was to design and implement a locally deployable virtual physician system, referred to as MedChat, capable of operating entirely offline within clinical environments. The system is specifically developed to function without reliance on external servers or internet connectivity, thereby ensuring compliance with strict data protection regulations and maintaining the highest possible standards of patient privacy and information security.

The MedChat system is conceptualized as an interactive virtual physician embedded within a controlled physical environment. In a typical clinical workflow, the patient enters a private, designated consultation room equipped with a high-resolution display, camera, and computing interface. Within this space, the virtual physician is projected on the screen, enabling a naturalistic and confidential interaction. Prior to initiating the consultation, the patient is registered by clinical staff, and relevant portions of the patient's medical history are securely preloaded into the system's local database. This allows the virtual physician to access contextually relevant information during the anamnesis process while preserving data locality and avoiding external data transmission.

During the consultation, MedChat autonomously conducts a structured anamnesis, guiding the patient through a series of context-aware questions designed to capture essential clinical information. The collected data are then processed, organized, and securely stored in a local database using encrypted storage protocols. Following each session, the system automatically generates a structured summary report that synthesizes the key findings of the interaction. This report can subsequently be accessed and reviewed by licensed healthcare professionals, providing an efficient and standardized foundation for further clinical assessment and decision-making.

Figure 1 presents a schematic overview of the anamnesis workflow facilitated by the MedChat system. By partially automating the patient intake procedure, MedChat contributes to the standardization of anamnesis across healthcare facilities, enhances the consistency of data collection, and optimizes the utilization of patient waiting times. Moreover, by reducing the administrative burden on medical personnel, the system allows healthcare providers to allocate more time to direct patient care and complex diagnostic reasoning. In this way, MedChat serves as both a technological innovation and a workflow enhancement tool, bridging artificial intelligence capabilities with the practical requirements of clinical practice.

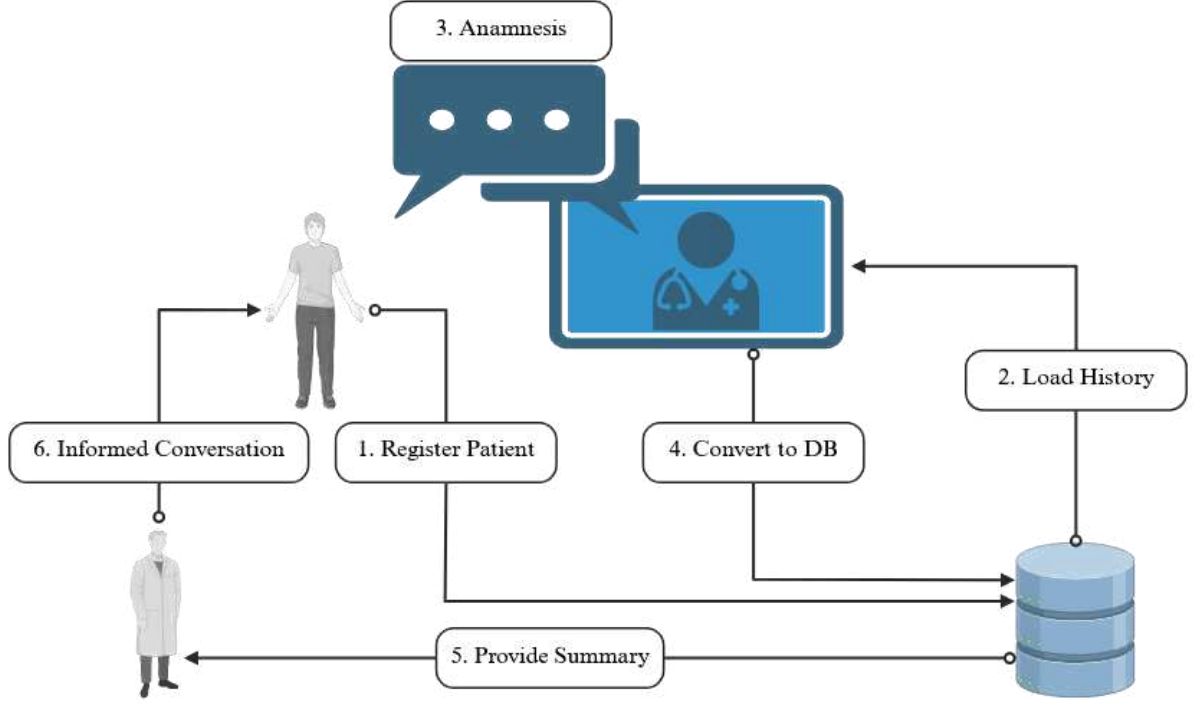


Figure 1: Overview of the anamnesis process facilitated by the MedChat system, illustrating its integration within standard clinical workflows.

2.2 The Virtual Doctor Avatar

2.2.1. Network Design and Preprocessing

We developed a comprehensive preprocessing pipeline to extract both audio and video frames from recorded materials. The audio signals were resampled to 22,050 Hz to enhance the fine-tuning performance of the text-to-speech (TTS) model. Prior to transformation, the recordings were downsampled to 16 kHz to ensure compatibility with the Wav2Vec model [25], which was employed for automated transcription. Pretrained Wav2Vec models from [26] were utilized to generate bilingual transcriptions in English and German. Subsequently, the audio data were converted into mel-frequency spectrograms using a window size of 1,024 and a hop length of 128, extracting 80 mel-frequency components to capture perceptually salient acoustic features.

The speech synthesis component was implemented using the Piper TTS framework [27], which follows an adversarial variational design inspired by the HiFi-GAN [28] and VITS [29] architectures. The training objective combines several complementary loss terms: (i) an adversarial loss applied to a least-squares discriminator-generator pair, (ii) a feature (perceptual) loss that minimizes the L1 distance between intermediate feature maps of real and generated audio, thereby enhancing perceptual fidelity in the frequency domain, and (iii) a Kullback-Leibler (KL) divergence term that regularizes the latent variable distribution. Together, these losses encourage both realistic waveform synthesis and consistent latent space structure, yielding speech output that closely resembles the target speaker’s tone and prosody. For this purpose, we adapted the pretrained Thorsten Voice model [30] for German and the Amy model [31] for English, training both on voice recordings provided by two researchers. The resulting customized voices ensured natural and consistent speech output in both languages, closely resembling the speaker’s tone and prosody.

To efficiently encode visual information, we developed a custom autoencoder that compresses image data into a compact latent-space representation, thereby reducing the computational cost of subsequent video generation. The model architecture was inspired by the Stable Diffusion framework [32] and trained on facial image datasets supplied by the two researchers with video recordings around 30 minutes yielding over 70,000 images each with corresponding audio signals. For initial experiments, we additionally employed the “Visual Historical Record of American High School Yearbooks” dataset [33] to assess model generalization. The latent space was constrained to the range $[-1, 1]$ using a hyperbolic tangent activation function to enhance numerical stability. Although this constraint may distort direct decoding, it improves robustness to noise and supports high-quality image reconstruction. To mitigate potential distortion, an adaptive normalization layer [34] was integrated into the decoder to map the constrained latent space back into an unconstrained representation. Furthermore, noise augmentation was introduced into the latent space to simulate imperfections typical of diffusion processes, promoting resilience during image generation. The autoencoder was designed to operate independently of input resolution, maintaining proportional compression relative to the image dimensions. Details of the model architecture and experimental results supporting these design choices are provided in appendix A.

For the video generation stage, a conditional diffusion model was implemented, conditioned jointly on the input image and the mel-frequency representation of the corresponding audio frame. This multimodal conditioning enables synchronized lip movement and facial dynamics consistent with the spoken content. The diffusion model adopts a U-Net backbone architecture augmented with cross-attention layers, conditional embeddings, and multi-head self-attention mechanisms, ensuring both temporal coherence across sequential frames and semantic consistency between visual and auditory modalities.

During inference, the model diverges from conventional diffusion frameworks that typically initiate the generation process from random Gaussian noise. Instead, it begins from a pre-diffused default image, which serves as a structured prior, substantially reducing the number of denoising steps required for image reconstruction. This modification results in a significant decrease in computational cost and inference latency while maintaining high visual fidelity in the synthesized video output.

Furthermore, the system was evaluated for its ability to generate multiple subsequent frames, from the same pre-diffused input. This approach enables parallel frame generation, potentially bridging the performance gap between diffusion-based synthesis and real-time video rendering. The resulting architecture demonstrates that, with appropriate conditioning and initialization strategies, diffusion models can achieve interactive responsiveness suitable for real-time avatar-based communication systems.

Figure 2 illustrates the training process of the diffusion model, highlighting the key modifications introduced during inference. During training, the model receives as input the current frame, the mel-frequency representation of the corresponding audio, and the previous frame, in addition to the timestep provided by the cosine noise scheduler. The target frame is progressively diffused, and the network is trained to predict the added noise, enabling it to learn the denoising dynamics. During inference, the trained model generates a sequence of frames, corresponding to one second of video (30 frames), by conditioning on a fixed reference image and a weakly diffused version of the same image. The mel-frequency inputs are updated continuously to reflect the evolving audio, allowing the system to produce temporally coherent video frames that are synchronized with the speech signal. This approach ensures both high-fidelity reconstruction and real-time responsiveness in the Virtual Doctor Avatar.

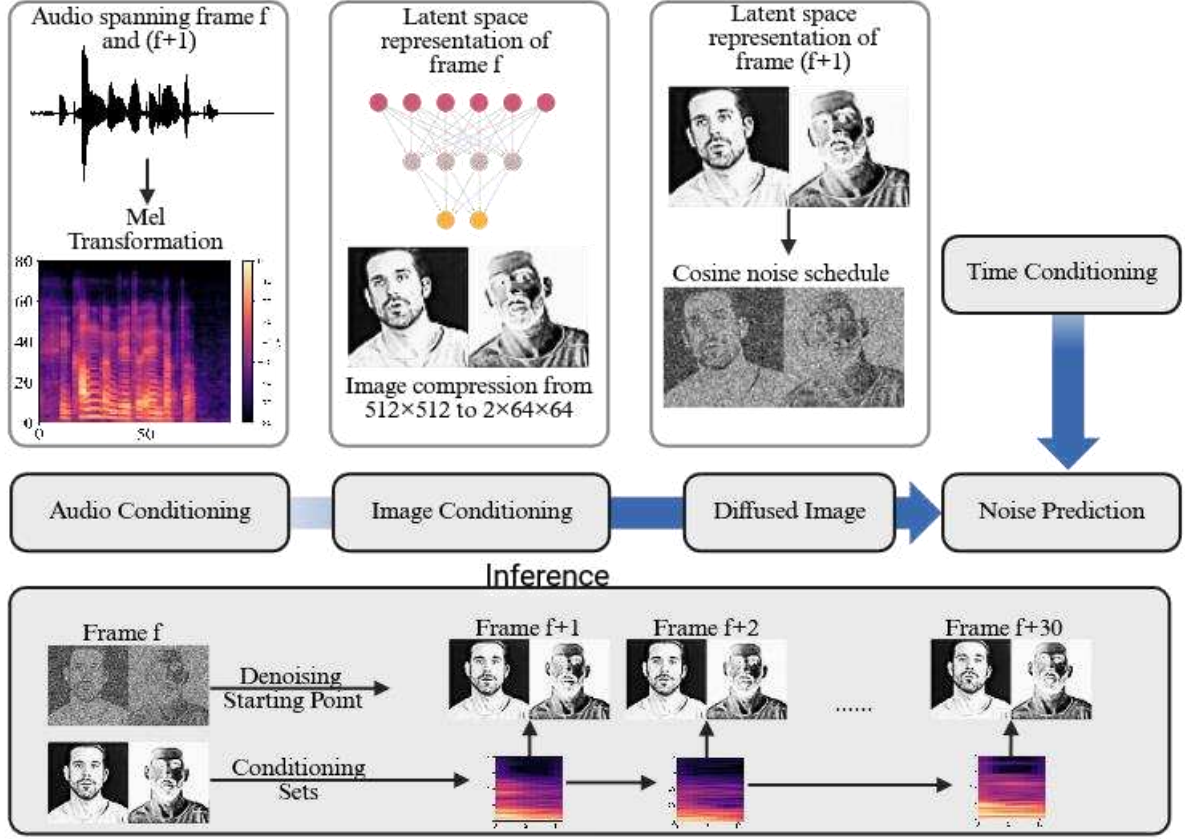


Figure 2: Illustration of the diffusion model training and inference pipeline for the Virtual Doctor Avatar. During training, the model is conditioned on mel-frequency features, the previous frame, and the current timestep from the cosine noise scheduler. The target frame is progressively diffused and used to train the network to predict the added noise. During inference, the trained model generates one second of video (30 frames) using a consistent conditioning image and a weakly diffused version of that image, while the conditioning mel-frequency inputs vary over time to reflect the changing audio features.

Table 2: Overview of model architectures and hyperparameters.

Component	Architecture	Loss Function	Epochs	Learning Rate
Autoencoder	Encoder-Decoder CNN	L1 Reconstruction	30	0.00001
Diffusion U-Net	Conditional U-Net with cross-attention	MSE	1,000	0.0001
Piper TTS	Variational Inference GAN	Adversarial, L1, KL	135	0.00001
MedChat (LLM)	LLaMA 3.1-8B with LoRA	Cross-Entropy	4	0.000005

2.2.2. Training Procedure

Having established the preprocessing and network design, we next detail the fine-tuning procedures applied to each component, focusing on training stability, convergence behavior, and computational efficiency. The training process commenced with the fine-tuning of the audio data using the Piper TTS model. A reduced learning rate of 0.00001 was applied to both the generator and discriminator components to promote stable convergence and prevent overfitting. Training was conducted over 135 epochs, with model checkpoints evaluated every 50 epochs to assess qualitative and quantitative improvements. After 135 epochs, the synthesized voice demonstrated a high degree of perceptual similarity to the target speaker. However, additional training beyond this point led to a gradual degradation in voice quality, suggesting the onset of overfitting and confirming the optimal stopping point for fine-tuning.

Following the completion of the speech synthesis stage, training proceeded to the video generation component, utilizing the previously developed autoencoder as the foundational model for latent space compression. After verifying the model’s robustness on the public image dataset [33] (see appendix A), the network was fine-tuned on recordings of the target researcher for 30 epochs. Training employed the Adam optimizer with an initial learning rate of 0.00001 and an L1 reconstruction loss as the objective function. A batch size of 32 was used, with all input images standardized to a resolution of 512×512 pixels. The autoencoder’s latent space was configured for a compression factor of 32.

For the diffusion-based generation stage, a cosine noise scheduling strategy [35] was employed to ensure smooth and stable convergence across 600 diffusion time steps. The model was trained for 1,000 epochs using the Adam optimizer with an initial learning rate of 0.0001, modulated by a cosine annealing schedule to facilitate adaptive step-size decay. The mean squared error (MSE) between the predicted and actual noise served as the training objective, consistent with standard Denoising Diffusion Probabilistic Model formulations [32].

2.3 MedChat Model Development

2.3.1 Datasets and Conversation Synthesis

To construct the primary dataset for large language model fine-tuning, we integrated three publicly available datasets from the Kaggle machine learning platform [36, 37, 38]. These datasets collectively comprised 10,080 samples, each containing a set of symptoms associated with a specific disease.

To generate realistic medical dialogues from these structured data samples, we employed a teacher-student knowledge distillation approach. Specifically, we fine-tuned the Meta-Llama-3.1-70B-Instruct model [39] to serve as a teacher model capable of generating high-quality, medically relevant conversations. The generated dialogues were then used to train a smaller, more deployment-friendly model, Meta-Llama-3.1-8B-Instruct, which forms the basis of our MedChat system.

We leveraged the capabilities of the larger, pretrained teacher model to reduce training time and benefit from its existing contextual understanding. However, due to the substantial hardware requirements (approximately 140 GB of VRAM in 16-bit precision) the 70B model is impractical for deployment in typical medical facility settings. Hence, our final MedChat implementation is based on the 8B version, which is more feasible for real-world applications.

To ensure the generated dialogues were clinically meaningful and structurally consistent, we provided the teacher model with detailed generation guidelines. These guidelines, described in the appendix B, included instructions on conversational tone, question structure, and appropriate responses to patient input. The model was allowed to generate up to 1,500 tokens per dialogue, enabling the creation of rich and informative medical conversations.

By repeating the generation process for every entry in the symptom-disease dataset, we created a large synthetic corpus of anamnesis dialogues. This synthetic dataset served as the foundation for fine-tuning our MedChat model, allowing it to effectively simulate medical interviews despite the absence of real patient interaction data.

Following the generation of 10,080 medical dialogues, we applied preprocessing steps to prepare the dialogues for model training, including formatting the conversations into structured JSON files. One common issue observed in the generated text was the excessive use of filler or acknowledgment phrases, e.g., “I see.” or “Noted”. To improve quality and coherence, we filtered out chatbot responses that did not contain a question and hence contained redundant information.

For training, 10% of the dataset was randomly withheld to serve as a validation set, enabling the assessment of overfitting and evaluation of model performance on an independent subset.

2.3.2 MedChat Training

We selected the Meta-Llama-3.1-8B-Instruct model [39] as the base model for fine-tuning, due to the impracticality of training a large language model from scratch, which would require extensive computational resources. Released in July 2024, this model is among the first robust large language models to combine a 128,000-token context window with multilingual capabilities, all within a manageable size of 8 billion parameters. Its relatively compact architecture allows it to be deployed and fine-tuned on consumer-grade or mid-range data center GPUs.

Fine-tuning was performed on a cluster of six NVIDIA A100 GPUs, each equipped with 40 GB of VRAM. We employed model parallelism by distributing instances of the model across all GPUs, splitting the dataset accordingly, and synchronizing gradients across devices during training. This setup enabled efficient parallel training and supported rapid experimentation. The dataset was divided into a 90:10 train-test split, with model evaluation conducted at the end of each epoch.

We used the AdamW optimizer [40] with a weight decay of 0.01 and an initial learning rate of 0.000005, coupled with a linear learning rate decay schedule. Each training batch consisted of 4 samples, and gradient accumulation was performed every 16 steps before synchronization across GPUs. Each epoch included 420 steps per GPU, and training was conducted for 4 epochs, resulting in approximately 1,680 training steps per device. In aggregate, this yielded 10,080 steps across all devices, corresponding to full coverage of our dataset.

To further enhance training efficiency and preserve the model’s broad capabilities, we adopted a parameter-efficient fine-tuning approach using Low-Rank Adaptation (LoRA) [41]. Since the model is required to retain general-purpose functionalities, such as open-domain conversation, classification, summarization, and multilingual support, we aimed to avoid catastrophic forgetting. In line with the recommendations from [41], we limited fine-tuning to lightweight adapter modules inserted into the Key and Value projection matrices of the transformer layers. This selective update approach provides

a favorable trade-off between computational cost and fine-tuning effectiveness while maintaining the integrity of the pretrained model's diverse skill set.

2.3.3 Database Integration

To manage and store data collected during interactions, we integrated an SQLite database into the MedChat system using SQLAlchemy, a Python-based Object-Relational Mapping framework. This architecture supports efficient and structured data handling, enabling seamless communication between the application and the underlying database layer.

Following the conclusion of each anamnesis session, MedChat produces a structured summary and classification of the dialogue in JSON format. This output is subsequently parsed, converted into application-level data objects, and persisted in the SQLite database. These operations (parsing, transformation, and data storage) are implemented outside the model itself and are managed by a dedicated service class, which functions as an intermediary between the model's output and the database. Although modern language models support direct database access via tool calls [39], we explicitly exclude this functionality to mitigate potential security risks and maintain strict control over data access.

To further safeguard the system against adversarial inputs, such as prompt injection and jailbreak attacks [42], we integrated Prompt Guard [43] into the application. Prompt Guard is a security-focused classification model that evaluates incoming prompts and categorizes them as benign, prompt injection, or jailbreak attempts. By embedding Prompt Guard into the input validation pipeline, we establish a proactive security layer that filters malicious content before it reaches the language model. This approach enhances the safety, reliability, and integrity of MedChat, particularly in the context of healthcare, where the consequences of model misuse can be significant.

3. Results

3.1 Training of the Autoencoder and Diffusion Model

The fine-tuning of the autoencoders, designed with a constrained latent space and augmented with latent-space noise, demonstrated stable and consistent convergence throughout the training process. The autoencoder was personalized using recordings from a single researcher, enabling the network to adapt to specific facial and vocal characteristics. Figure 3 presents the L1 reconstruction loss over the course of 30 training epochs. At the beginning of training, the L1 loss started at approximately 0.382 and rapidly decreased during the first 10 epochs, reaching a value of 0.051. The subsequent epochs exhibited a slower, steady decline, ultimately stabilizing around 0.037 by the final epoch. This smooth convergence indicates that the network successfully learned to encode and reconstruct input images while maintaining the constraints imposed on the latent space. The latent-space noise augmentation further contributed to this stability by preventing overfitting and promoting robustness to perturbations in the latent representation. Notably, the short fine-tuning duration of 30 epochs was sufficient to achieve a low reconstruction error, demonstrating the efficiency of the proposed architecture and training regimen.

The U-Net model employed for diffusion-based video generation also exhibited clear convergence over its training regime. Figure 4 illustrates the MSE value between the predicted and applied noise across

1,000 epochs. The training curve begins at an initial MSE of approximately 0.774 and shows a consistent downward trend, reaching around 0.016 by the final epochs. While minor fluctuations of up to ± 0.002 are observed throughout training, these are attributable to the stochastic nature of the diffusion process, where noise is randomly sampled at each timestep, and do not indicate instability. The consistent overall decrease in MSE demonstrates that the network effectively learned to denoise latent representations while conditioning on both image and audio features. These fluctuations also suggest that the model maintains adaptability to varying noise patterns, which is beneficial for generalization during inference.

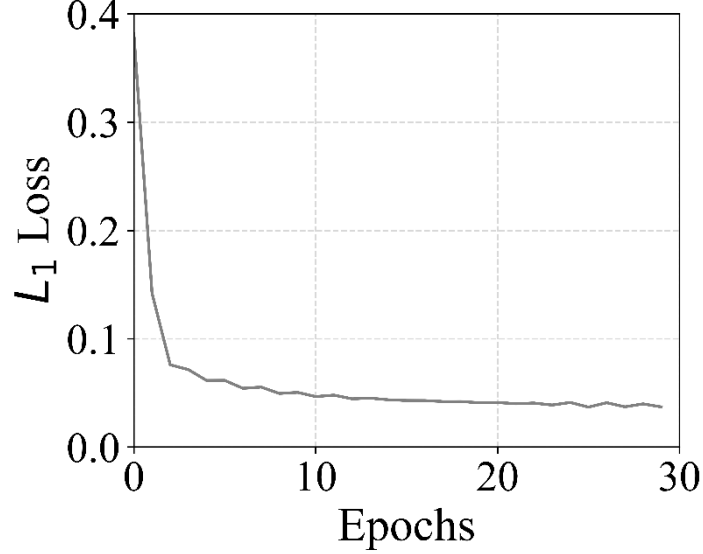


Figure 3: L1 reconstruction loss of the autoencoder during fine-tuning on recordings from a single researcher. The network was trained for 30 epochs with latent-space noise augmentation and a constrained latent representation. The curve demonstrates smooth and stable convergence, indicating effective learning of the input image reconstruction while maintaining latent-space regularity.

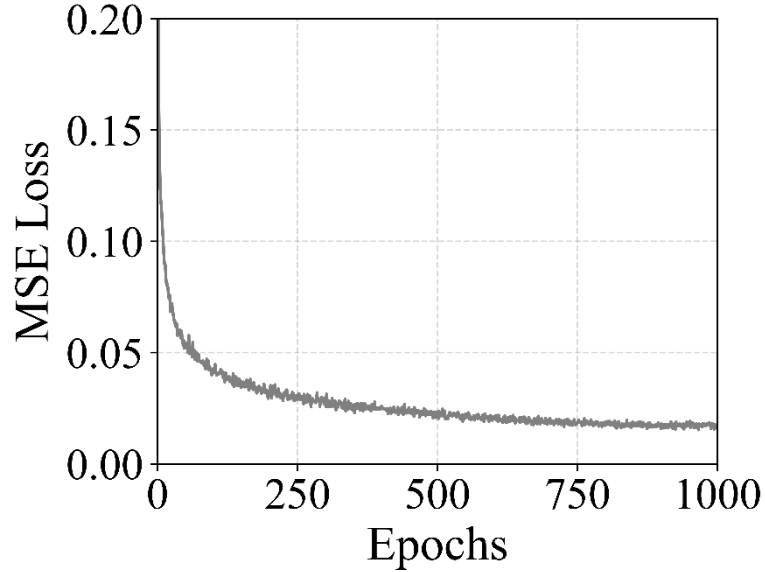


Figure 4: MSE value of the U-Net diffusion model during training over 1,000 epochs. The model predicts the noise added to latent representations conditioned on image and audio features using a cosine

noise schedule. Minor fluctuations are observed due to the stochastic nature of the diffusion process, while the overall downward trend demonstrates stable convergence.

3.2 Dialogue Generation and MedChat Fine-tuning

Figure 5 presents the fine-tuning dynamics of the MedChat model. At the onset of training, both the training and evaluation losses are relatively high, reflecting the model’s initial lack of adaptation to the specific task and substantial deviation from the desired output distributions. During the first 500 training steps, a rapid decrease in both losses is observed, indicating that the model quickly begins to capture relevant patterns from the training data. Beyond this point, the training loss continues to decrease gradually, while the evaluation loss stabilizes, demonstrating that the model maintains consistent performance on unseen data. The close alignment between training and evaluation loss throughout the remainder of the fine-tuning process suggests minimal overfitting and highlights the model’s ability to generalize effectively. Overall, these results indicate that the fine-tuning procedure successfully adapts MedChat to the target task while preserving robustness to novel inputs.

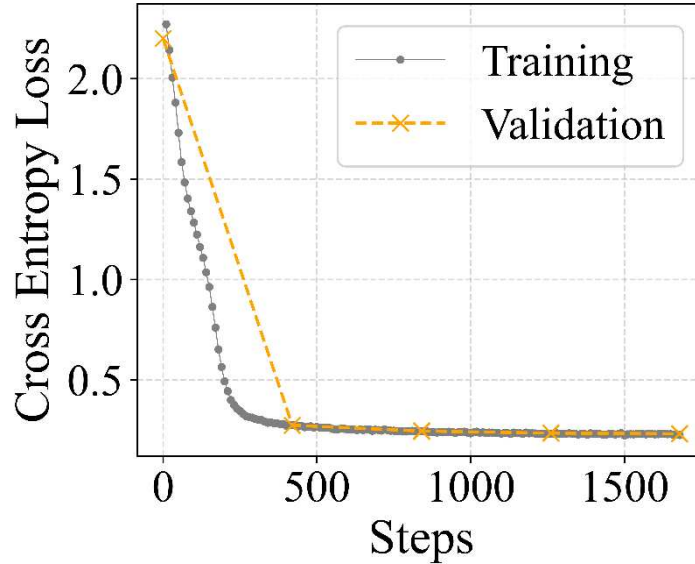


Figure 5: Training loss curves for the MedChat fine-tuning process. The model exhibits a rapid initial decrease in loss, followed by convergence, indicating effective learning and stable generalization to unseen data.

The simulated interaction described in Appendix C involves an initial consultation with a patient who has no prior history recorded in the MedChat system and for whom no external information from other models or sensor data is available. In this simulation, the patient is modeled as presenting symptoms consistent with pneumonia. The scenario is designed to evaluate the ability of MedChat to autonomously conduct a structured anamnesis, elicit relevant clinical information, and generate an initial patient summary entirely based on the interaction within the system. This setup provides a controlled environment for assessing the model’s performance in handling cases without any pre-existing contextual knowledge or external guidance.

From a deployment perspective, the full MedChat-Avatar system can operate on consumer-grade GPUs with as little as 40 GB of VRAM for end-to-end interaction, or on 24 GB of VRAM and only CPUs when the audio-only mode is used. This configuration substantially reduces hardware costs and energy consumption compared to cloud-based LLM solutions, which typically require 80-100 GB VRAM or distributed clusters for inference.

Table 3 provides a short summary of the models developed in the research as well as their training behavior.

Table 3: Quantitative summary of training performance across components.

Model	Initial Loss	Final Loss	Epochs	Convergence Behavior
Autoencoder	0.382	0.037	30	Smooth convergence
Diffusion Model	0.774	0.016	1,000	Stable with minor stochastic fluctuations
MedChat	2.45	0.23	4	Stable convergence without overfitting

4. Discussion

The developed medical chatbot, MedChat, is designed as a clinical decision support tool rather than a stand-alone diagnostic system. All interactions require supervision by qualified medical personnel, and no direct diagnostic or therapeutic functions are performed autonomously. Despite these constraints, the system demonstrates robust performance under realistic hardware conditions, validating the feasibility of offline deployment.

From a deployment perspective, MedChat is computationally accessible, as the base model can be run on consumer-grade GPUs equipped with at least 24 GB of VRAM, making it feasible for integration into local healthcare infrastructure without high-end computational resources. In addition, the Virtual Doctor avatar interface can run on 16 GB of VRAM or a reduced version where only audio is available can run without a graphic card and is even manageable by raspberry Pi 4b.

The current implementation of the avatar operates under a Markovian assumption, wherein each generated frame is conditioned solely on the preceding frame and the most recent audio segment. While this design choice significantly reduces computational complexity and enables efficient real-time generation, it also introduces limitations in temporal coherence. Specifically, the resulting motion between consecutive frames may appear subtle or slightly unnatural due to the absence of long-term temporal dependencies. Future improvements could involve extending the conditioning context through a 3D convolutional U-Net architecture [44], enabling the model to process and generate entire video sequences with awareness of multiple preceding and future frames. Such an approach would likely enhance temporal smoothness and realism in facial and gestural movements and thereby overcoming the uncanny valley. However, these benefits would come at the expense of increased computational requirements during both training and inference. Thus, the Markovian framework currently offers a pragmatic trade-off between performance efficiency and visual naturalness, serving as a foundation for subsequent advancements in temporally coherent avatar synthesis.

In terms of performance evaluation, the Virtual Doctor avatar, comprising the audio and image generation components, was qualitatively assessed using L1 loss to ensure fidelity to the expected latent space representations. However, such quantitative evaluation could not be directly applied to the

MedChat dialogue model, due to the inherent subjectivity and open-ended nature of conversational interaction. Instead, the performance of MedChat will be systematically assessed through an upcoming study at the University Hospital Bochum, where its real-world applicability and usability will be evaluated under medical supervision.

Preliminary results, included in the appendix C, illustrate natural and coherent dialogue between the avatar and users, suggesting promising potential for interactive anamnesis. Nonetheless, several limitations must be acknowledged.

First, the training data for MedChat includes synthetically generated dialogues and relies heavily on the pretrained knowledge of the base LLaMA model [39], which may embed inherent biases. These biases could lead to unexpected behaviors or misinterpretations during the anamnesis process, particularly in complex or ambiguous clinical cases. It should always be used with care and medical supervision.

Second, the model’s ability to generalize to rare symptoms or conditions is limited by the coverage and diversity of the synthetic dataset. As a result, MedChat may underperform when faced with less common medical presentations not adequately represented in the training data. Thus, it has to be improved in the future with corrected anamnesis results provided by physicians.

Third, the model’s medical knowledge base is restricted to the information contained within its training corpus. This constraint implies that MedChat may lack awareness of emerging diseases, new clinical guidelines, or the latest evidence-based practices. This gap could compromise the accuracy and completeness of the anamnesis it conducts, highlighting the need for continued fine-tuning using large-scale, up-to-date medical corpora curated from verified sources.

In summary, MedChat represents a significant advancement in the development of AI-driven clinical support systems, combining natural language dialogue with a multimodal virtual avatar capable of generating realistic audio and visual outputs. This integration enhances the user experience by creating a more engaging and human-like interaction during the anamnesis process. While the current system has limitations related to training data coverage, potential biases, and the need for clinical validation, it already demonstrates the ability to conduct structured, medically relevant conversations in a natural and intuitive manner.

Importantly, MedChat is designed to run on accessible, consumer-grade hardware and follows a modular architecture, making it adaptable for a range of clinical environments. The forthcoming clinical evaluation and further fine-tuning on large-scale, domain-specific medical corpora are expected to improve its accuracy, generalizability, and safety. With continued development, MedChat, together with its Virtual Doctor avatar, has the potential to become a valuable tool in supporting healthcare professionals, streamlining patient interviews, and improving access to consistent, high-quality preliminary assessments. This work lays a strong foundation for the next generation of intelligent, multimodal healthcare assistants, capable of complementing human expertise and enriching patient care.

Future work will extend the temporal coherence of the avatar via 3D convolutional architectures and expand MedChat’s conversational robustness through reinforcement learning with human feedback. The planned clinical evaluation at the University Hospital Bochum will further quantify usability, patient trust, and integration feasibility in real-world workflows.

5. Conclusion

In this paper, we present a fully AI-driven anamnesis pipeline designed to autonomously conduct patient interviews and generate concise, structured summaries to assist physicians by highlighting key medical information. The system, referred to as MedChat, integrates language models optimized for local execution, enabling deployment on consumer-grade graphics hardware. This ensures accessibility and practicality for clinical environments with limited computational infrastructure.

MedChat was developed using a combination of medical datasets and synthetically generated dialogues. However, it is important to note that the system has not yet undergone formal clinical validation. A prospective evaluation is planned at the University Hospital in Bochum, where the system's utility will be assessed from two perspectives: the extent to which it supports medical professionals in their diagnostic workflow, and the degree of acceptance and trust exhibited by patients toward an AI-guided anamnesis process.

Future development efforts will focus on integrating more advanced LLMs and enhancing the visual and interactive quality of the system's virtual avatar. A key ongoing challenge is maintaining a balance between system performance and practical feasibility. Many medical institutions operate under significant budget constraints and lack the resources required to deploy large-scale AI models. As such, our design prioritizes computational efficiency to offer a lightweight and cost-effective solution without compromising core functionality.

To promote transparency and foster further research in the field, we release both the synthetic training data and the developed neural network architectures as open-source software. This enables reproducibility of our results and provides a standardized benchmark for the development of future medical chatbot systems.

The system will be integrated into the Virtual Doctor project with testing and medical evaluation performed at the University Hospital Bochum in Germany. The results will be used to further improve on the framework and increase the benchmark dataset.

Ethics Statement

All experiments were performed using synthetic or publicly available datasets. No human or patient data were collected. All procedures comply with institutional and national ethical guidelines, including the EU General Data Protection Regulation (GDPR).

Author Contributions

D. H. served as the principal investigator and was responsible for overseeing the conceptualization and implementation of all components. J. B. R. developed the virtual avatar pipeline and supervised the development of the MedChat software. D. B. led the conceptualization and core development of the MedChat system. J. P. S. and A. C. contributed to the regulatory preparations necessary for future clinical testing of MedChat at the University Hospital Bochum, Germany.

Funding

Not available.

5. Conclusion

In this paper, we present a fully AI-driven anamnesis pipeline designed to autonomously conduct patient interviews and generate concise, structured summaries to assist physicians by highlighting key medical information. The system, referred to as MedChat, integrates language models optimized for local execution, enabling deployment on consumer-grade graphics hardware. This ensures accessibility and practicality for clinical environments with limited computational infrastructure.

MedChat was developed using a combination of medical datasets and synthetically generated dialogues. However, it is important to note that the system has not yet undergone formal clinical validation. A prospective evaluation is planned at the University Hospital in Bochum, where the system's utility will be assessed from two perspectives: the extent to which it supports medical professionals in their diagnostic workflow, and the degree of acceptance and trust exhibited by patients toward an AI-guided anamnesis process.

Future development efforts will focus on integrating more advanced LLMs and enhancing the visual and interactive quality of the system's virtual avatar. A key ongoing challenge is maintaining a balance between system performance and practical feasibility. Many medical institutions operate under significant budget constraints and lack the resources required to deploy large-scale AI models. As such, our design prioritizes computational efficiency to offer a lightweight and cost-effective solution without compromising core functionality.

To promote transparency and foster further research in the field, we release both the synthetic training data and the developed neural network architectures as open-source software. This enables reproducibility of our results and provides a standardized benchmark for the development of future medical chatbot systems.

The system will be integrated into the Virtual Doctor project with testing and medical evaluation performed at the University Hospital Bochum in Germany. The results will be used to further improve on the framework and increase the benchmark dataset.

Ethics Statement

All experiments were performed using synthetic or publicly available datasets. No human or patient data were collected. All procedures comply with institutional and national ethical guidelines, including the EU General Data Protection Regulation (GDPR).

Author Contributions

D. H. served as the principal investigator and was responsible for overseeing the conceptualization and implementation of all components. J. B. R. developed the virtual avatar pipeline and supervised the development of the MedChat software. D. B. led the conceptualization and core development of the MedChat system. J. P. S. and A. C. contributed to the regulatory preparations necessary for future clinical testing of MedChat at the University Hospital Bochum, Germany.

Funding

Not available.

Data Availability Statement

Due to ethical and privacy considerations, the trained autoencoder and diffusion model weights are not openly released but can be made available upon reasonable research request and subject to appropriate data use agreements.

References

- [1] Jarrahi, Mohammad Hossein. (2018). Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. *Business Horizons*. 61. 10.1016/j.bushor.2018.03.007.
- [2] Sarumi, Oluwafemi & Heider, Dominik. (2024). Large language models and their applications in bioinformatics. *Computational and Structural Biotechnology Journal*. 23. 10.1016/j.csbj.2024.09.031.
- [3] Shi, Yi & Sun, Lin. (2024). How Generative AI Is Transforming Journalism: Development, Application and Ethics. *Journalism and Media*. 5. 582-594. 10.3390/journalmedia5020039.
- [4] Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, Harari YN, Zhang YQ, Xue L, Shalev-Shwartz S, Hadfield G, Clune J, Maharaj T, Hutter F, Baydin AG, McIlraith S, Gao Q, Acharya A, Krueger D, Dragan A, Torr P, Russell S, Kahneman D, Brauner J, Mindermann S. Managing extreme AI risks amid rapid progress. *Science*. 2024 May 24;384(6698):842-845. doi: 10.1126/science.adn0117. Epub 2024 May 20. PMID: 38768279.
- [5] Alberts IL, Mercolli L, Pyka T, Prenosil G, Shi K, Rominger A, Afshar-Oromieh A. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *Eur J Nucl Med Mol Imaging*. 2023 May;50(6):1549-1552. doi: 10.1007/s00259-023-06172-w. Epub 2023 Mar 9. PMID: 36892666; PMCID: PMC9995718.
- [6] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI Blog.
- [7] Obradovich N, Khalsa SS, Khan W, Suh J, Perlis RH, Ajilore O, Paulus MP. Opportunities and Risks of Large Language Models in Psychiatry. *NPP Digit Psychiatry Neurosci*. 2024;2(1):8. doi: 10.1038/s44277-024-00010-z. Epub 2024 May 24. PMID: 39554888; PMCID: PMC11566298.
- [8] Morley J, Machado CCV, Burr C, Cows J, Joshi I, Taddeo M, Floridi L. The ethics of AI in health care: A mapping review. *Soc Sci Med*. 2020 Sep;260:113172. doi: 10.1016/j.socscimed.2020.113172. Epub 2020 Jul 15. PMID: 32702587.
- [9] Tajabadi, Mohammad & Martin, Roman & Heider, Dominik. (2024). Privacy-Preserving Decentralized Learning Methods for Biomedical Applications. *Computational and Structural Biotechnology Journal*. 23. 10.1016/j.csbj.2024.08.024.
- [10] Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, Uhlmann L; Reader study level-I and level-II Groups; Alt C, Arenbergerova M, Bakos R, Baltzer A, Bertlich I, Blum A, Bokor-Billmann T, Bowling J, Braghiroli N, Braun R, Buder-Bakhaya K, Buhl T, Cabo H, Cabrijan L, Cevic N, Classen A, Deltgen D, Fink C, Georgieva I, Hakim-Meibodi LE, Hanner S, Hartmann F, Hartmann J, Haus G, Hoxha E, Karls R, Koga H, Kreusch J, Lallas A, Majenka P, Marghoob A, Massone C, Mekokishvili L, Mestel D, Meyer V,

Neuberger A, Nielsen K, Oliviero M, Pampena R, Paoli J, Pawlik E, Rao B, Rendon A, Russo T, Sadek A, Samhaber K, Schneiderbauer R, Schweizer A, Toberer F, Trennheuser L, Vlahova L, Wald A, Winkler J, Wölbing P, Zalaudek I. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018 Aug 1;29(8):1836-1842. doi: 10.1093/annonc/mdy166. PMID: 29846502.

[11] Choi YJ, Park MJ, Ko Y, Soh MS, Kim HM, Kim CH, Lee E, Kim J. Artificial intelligence versus physicians on interpretation of printed ECG images: Diagnostic performance of ST-elevation myocardial infarction on electrocardiography. *Int J Cardiol*. 2022 Sep 15;363:6-10. doi: 10.1016/j.ijcard.2022.06.012. Epub 2022 Jun 9. PMID: 35691440.

[12] Gao Z, Xu Y, Sun C, Wang X, Guo Y, Qiu S, Ma K. A systematic review of asymptomatic infections with COVID-19. *J Microbiol Immunol Infect*. 2021 Feb;54(1):12-16. doi: 10.1016/j.jmii.2020.05.001. Epub 2020 May 15. PMID: 32425996; PMCID: PMC7227597.

[13] Tu, T., Schaekermann, M., Palepu, A. *et al*. Towards conversational diagnostic artificial intelligence. *Nature* (2025). <https://doi.org/10.1038/s41586-025-08866-7>

[14] Ran Xu, Wenqi Shi, Jonathan Wang, Jasmine Zhou, and Carl Yang. 2025. MedAssist: LLM-Empowered Medical Assistant for Assisting the Scrutinization and Comprehension of Electronic Health Records. In Companion Proceedings of the ACM on Web Conference 2025 (WWW '25). Association for Computing Machinery, New York, NY, USA, 2931–2934. <https://doi.org/10.1145/3701716.3715186>

[15] Lu, Meng & Ho, Brandon & Ren, Dennis & Wang, Xuan. (2024). TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage. 5747-5764. 10.18653/v1/2024.findings-emnlp.329.

[16] J V, Lavan & Sangeetha, Lakshmi. (2025). HealsHealthAI: Unveiling Personalized Healthcare Insights with Open Source Fine-Tuned LLM. 10.1002/9781394249312.ch4.

[17] Ruhland JB, Wichmann J, Degtyar D, Martin R, Fehse L, Klau JH, Papenbrock T, Sowa JP, Canbay A, Freisleben B, Leyer M, Heider D. The virtual doctor prescribing the future: Diagnostics with interactive clinical decision support. *Comput Biol Med*. 2025 Sep;196(Pt C):110968. doi: 10.1016/j.combiomed.2025.110968. Epub 2025 Aug 20. PMID: 40839937.

[18] Yingbing Huang, Lily Jiabin Wan, Hanchen Ye, Manvi Jha, Jinghua Wang, Yuhong Li, Xiaofan Zhang, and Deming Chen. 2024. Invited: New Solutions on LLM Acceleration, Optimization, and Application. In Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC '24). Association for Computing Machinery, New York, NY, USA, Article 369, 1–4. <https://doi.org/10.1145/3649329.3663517>

[19] Winkler A, Kutschar P, Pitzer S, van der Zee-Neuen A, Kerner S, Osterbrink J, Krutter S. Avatar and virtual agent-assisted telecare for patients in their homes: A scoping review. *J Telemed Telecare*. 2025 Feb;31(2):207-221. doi: 10.1177/1357633X231174484. Epub 2023 Jun 7. PMID: 37287248.

[20] DeTore, Nicole & Balogun, Oyenike & Eberlin, Elizabeth & Dokholyan, Katherine & Rizzo, Albert & Holt, Daphne. (2024). An artificial intelligence-based virtual human avatar application to

assess the mental health of healthcare professionals: A validation study (Preprint).
10.2196/preprints.58255.

[21] Goodfellow, I. J., “Generative Adversarial Networks”, arXiv:1406.2661, 2014.
doi:10.48550/arXiv.1406.2661.

[22] Ho, J., Jain, A., and Abbeel, P., “Denoising Diffusion Probabilistic Models”, Art. no. arXiv:2006.11239, 2020. doi:10.48550/arXiv.2006.11239.

[23] Deutch, G., Gal, R., Garibi, D., Patashnik, O., and Cohen-Or, D., “TurboEdit: Text-Based Image Editing Using Few-Step Diffusion Models”, <i>arXiv e-prints</i>, Art. no. arXiv:2408.00735, 2024. doi:10.48550/arXiv.2408.00735.

[24] Salimans, T. and Ho, J., “Progressive Distillation for Fast Sampling of Diffusion Models”, <i>arXiv e-prints</i>, Art. no. arXiv:2202.00512, 2022. doi:10.48550/arXiv.2202.00512.

[25] Baevski, A., Zhou, H., Mohamed, A., and Auli, M., “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”, <i>arXiv e-prints</i>, Art. no. arXiv:2006.11477, 2020. doi:10.48550/arXiv.2006.11477.

[26] J. Grosman, “Fine-tuned XLSR-53 large model for speech recognition in German,” Hugging Face, 2021. [Online]. Available: <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>

[27] OHF-Voice, “piper1-gpl: Fast and local neural text-to-speech engine,” Github. [Online]. Available: <https://github.com/OHF-Voice/piper1-gpl>

[28] Kong, J., Kim, J., and Bae, J., “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis”, arXiv:2010.05646, 2020. doi:10.48550/arXiv.2010.05646.

[29] Kim, J., Kong, J., and Son, J., “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech”, arXiv:2106.06103, 2021. doi:10.48550/arXiv.2106.06103.

[30] M. Hansen, “Thorsten medium checkpoint,” Rhasspy, 2023. [Dataset]. Available: https://huggingface.co/datasets/rhasspy/piper-checkpoints/tree/main/de/de_DE/thorsten/medium

[31] M. Hansen, “Amy medium checkpoint,” Rhasspy, 2023. [Dataset]. Available: https://huggingface.co/datasets/rhasspy/piper-checkpoints/tree/main/en/en_US/amy/medium

[32] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., “High-Resolution Image Synthesis with Latent Diffusion Models”, <i>arXiv e-prints</i>, Art. no. arXiv:2112.10752, 2021. doi:10.48550/arXiv.2112.10752.

[33] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, Crystal Lee, Philipp Krähenbühl, Alexei A. Efros. A Century of Portraits: A Visual Historical Record of American High School Yearbooks, Extreme Imaging Workshop, International Conference on Computer Vision, ICCV 2015. and IEEE Transactions on Computational Imaging, September 2017.

- [34] Jan Benedikt Ruhland, Iraj Masoudian, Dominik Heider, Enhancing deep neural network training through learnable adaptive normalization, Knowledge-Based Systems, Volume 326, 2025, 113968, ISSN 0950-7051, <https://doi.org/10.1016/j.knosys.2025.113968>.
- [35] Nichol, A. and Dhariwal, P., “Improved Denoising Diffusion Probabilistic Models”, <i>arXiv e-prints</i>, Art. no. arXiv:2102.09672, 2021. doi:10.48550/arXiv.2102.09672.
- [36] Disease prediction based on symptoms. en. url: <https://www.kaggle.com/datasets/pasindueranga/disease-prediction-based-on-symptoms> (retrieved on 28.10.2024).
- [37] Disease Prediction Using Machine Learning. url: <https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning?resource=download> (retrieved on 28. 10. 2024).
- [38] Disease and Symptoms dataset. en. url: <https://www.kaggle.com/datasets/choongqianzheng/disease-and-symptoms-dataset> (retrieved on 28. 10. 2024).
- [39] Abhimanyu Dubey u. a. The Llama 3 Herd of Models. arXiv:2407.21783. Aug. 2024. doi: 10.48550/arXiv.2407.21783. url: <http://arxiv.org/abs/2407.21783>
- [40] Ilya Loshchilov und Frank Hutter. Decoupled Weight Decay Regularization. arXiv:1711.05101. Jan. 2019. doi: 10.48550/arXiv .1711.05101. url: <http://arxiv.org/abs/1711.05101>
- [41] Edward J. Hu u. a. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685. Okt. 2021. doi: 10.48550/arXiv.2106.09685. url: <http://arxiv.org/abs/2106.09685>
- [42] Yi Liu u. a. Prompt Injection attack against LLM-integrated Applications. arXiv:2306.05499. March 2024. doi: 10 . 48550 / arXiv . 2306 . 05499. url: <http://arxiv.org/abs/2306.05499>
- [43] meta-llama/Prompt-Guard-86M · Hugging Face. Sep. 2024. url: <https://huggingface.co/meta-llama/Prompt-Guard-86M> (visited on 29. 10. 2024).
- [44] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J., “Video Diffusion Models”, arXiv:2204.03458, 2022. doi:10.48550/arXiv.2204.03458.

Appendix A. Ablation Experiments

Our initial experiments were conducted using the selected dataset to design an autoencoder with a stable and well-constrained latent space, thereby improving training robustness during the subsequent diffusion modeling stage. To assess the impact of different latent space regularization strategies, we compared three autoencoder configurations based on the design described in table A1.

Table A1: Overview of the autoencoder architecture developed for this study. The model was implemented and tested across three configurations to investigate the influence of latent space regularization. In configuration 1, the tanh activation function at the encoder output was omitted. Configuration 2 integrated the reparameterization trick [32] to enforce a probabilistic latent distribution. Configuration 3, described in detail in this table, constitutes the final configuration adopted for all subsequent diffusion model experiments.

Module	Layer / Operation	Configuration	Output Resolution	Notes
Encoder	Adaptive Normalization	AdaptiveNorm((1, in_channels, 1, 1))	$N \times N$	Input normalization
	Convolution	Conv2d(in_channels, 32, 3, 1, 1)	$N \times N$	Initial feature extraction
	ResBlock	ResBlock(32, 32)	$N \times N$	Residual feature refinement
	Downsampling	Conv2d(32, 32, 4, 2, 1)	$N/2 \times N/2$	First spatial downsampling
	ResBlock	ResBlock(32, 64)	$N/2 \times N/2$	Feature expansion
	Downsampling	Conv2d(64, 64, 4, 2, 1)	$N/4 \times N/4$	Second spatial downsampling
	ResBlock	ResBlock(64, 128)	$N/4 \times N/4$	Deep feature encoding

	Downsampling	Conv2d(128, 128, 4, 2, 1)	$N/8 \times N/8$	Latent compression
	ResBlock	ResBlock(128, 128)	$N/8 \times N/8$	High-level feature fusion
	Projection	Conv2d(128, latent_channels, 3, 1, 1)	$N/8 \times N/8$	Latent embedding
	Activation	Tanh()	$N/8 \times N/8$	Constrains latent space to $[-1, 1]$
Decoder	Adaptive Normalization	AdaptiveNorm((1, latent_channels, 1, 1))	$N/8 \times N/8$	Normalization of latent input
	Linear Projection	Conv2d(latent_channels, 128, 1)	$N/8 \times N/8$	Expands latent space into feature domain
	ResBlock	ResBlock(128, 128)	$N/8 \times N/8$	Residual learning in latent domain
	Upsampling	ConvTranspose2d(128, 64, 4, 2, 1)	$N/4 \times N/4$	First spatial upsampling
	ResBlock	ResBlock(64, 64)	$N/4 \times N/4$	Feature refinement
	Upsampling	ConvTranspose2d(64, 32, 4, 2, 1)	$N/2 \times N/2$	Second spatial upsampling
	ResBlock	ResBlock(32, 32)	$N/2 \times N/2$	Feature refinement

	Upsampling	ConvTranspose2d(32, 16, 4, 2, 1)	N×N	Final spatial upsampling
	ResBlock	ResBlock(16, 16)	N×N	Final feature smoothing
	Output Convolution	Conv2d(16, out_channels, 3, 1, 1)	N×N	Image reconstruction
	Activation	Tanh()	N×N	Constrains pixel values to [-1, 1]

Our first approach was to establish a base line with an unconstrained latent space, i.e. where the latent variables were allowed to take arbitrary continuous values. The second approach was a variationally constrained latent space, where a normal prior distribution $N(0, I)$ was enforced through the reparameterization trick and an additional Kullback-Leibler divergence loss term as described in [32]. The last approach contained a deterministically constrained latent space, where a hyperbolic tangent activation was applied to project latent variables into the range $[-1, 1]$. An adaptive normalization layer [34] was incorporated in the decoder to re-establish consistent statistics during reconstruction.

The reparameterization trick introduces stochasticity into the latent representation while maintaining differentiability, enabling efficient gradient-based optimization. It can be expressed as:

$$(1) \quad z = \mu + \sigma \times \epsilon, \epsilon \sim N(0, I)$$

where μ and σ represent the mean and standard deviation predicted by the encoder, and ϵ denotes a noise sample drawn from a standard normal distribution. To regularize the latent space toward a standard normal prior, we employed the Kullback-Leibler divergence loss, defined as:

$$(2) \quad L_{KL} = -0.5 \sum_{i=1}^d (1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2)$$

where the sum is over the full dimensionality d of the latent space.

The reconstruction objective for all configurations was the L1 loss between the input and reconstructed image:

$$(3) \quad L_{rec} = \|x - \hat{x}\|_1$$

The autoencoder models were trained for 30 epochs using images of 256×256 resolution. To enhance the generalization capability of the network, a set of geometric data augmentation techniques was applied, including random rotations, translations, scalings, and shear transformations. All transformations employed zero padding to preserve image dimensions. The training objective combined

an L1 reconstruction loss, as defined in equation (3), with a Kullback-Leibler divergence term serving as a regularization component for configuration 2. The regularization coefficient was varied between 0.000001 and 0.01 to assess its impact on latent space stability. Optimization was performed using the AdamW algorithm [38], which provides adaptive moment estimation with decoupled weight decay to improve convergence behavior.

Figure A6 presents the evolution of the L1 reconstruction loss across all three autoencoder configurations. The relatively weaker reconstruction performance observed in configuration 2 at a regularization weight of 0.01 can be attributed to the increased stochasticity introduced by the KL divergence, which induces variability in the decoding process. When the regularization coefficient was reduced to 0.000001, the reconstruction quality improved. However, the latent space exhibited poor constraint, with activation magnitudes exceeding an order of magnitude of 10. Balancing the KL and L1 loss terms proved challenging, as excessive regularization reduced reconstruction fidelity, whereas insufficient regularization led to latent instability. Consequently, configuration 3 was selected as the final model, as it achieved a well-defined latent representation suitable for subsequent latent space noise augmentation, thereby enhancing the robustness of the downstream diffusion model.

Figure A7 provides a qualitative comparison between the original input images and their corresponding reconstructions for each autoencoder configuration, illustrating the inherent trade-off between reconstruction fidelity and latent space regularity. The results for the strongly KL-regularized autoencoder are omitted, as training exhibited strong KL divergence dominance. Consequently, although the latent space distribution closely approximated a normal prior, the reconstruction quality degraded substantially. This behavior, consistent with the quantitative results shown in figure A6, indicates that excessive KL regularization overly constrains the latent representation, thereby impairing the network’s ability to accurately reconstruct fine-grained image details.

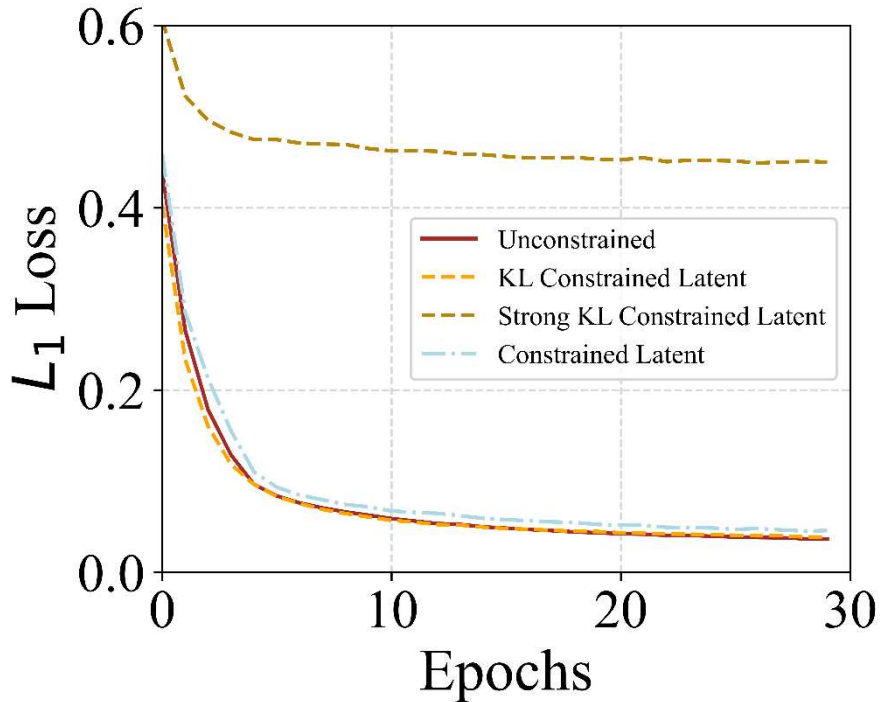


Figure A6: Training curves for the different autoencoder configurations. For configuration 2, which incorporates a KL divergence term, multiple regularization strengths were evaluated to assess the effect of relaxing the normal prior assumption. The curve labeled strong KL constraint corresponds to

a regularization coefficient of 0.01, reflecting a strong enforcement of the prior, whereas the weak KL constraint (orange curve) employs a coefficient of 0.000001, allowing greater flexibility in the latent space representation.

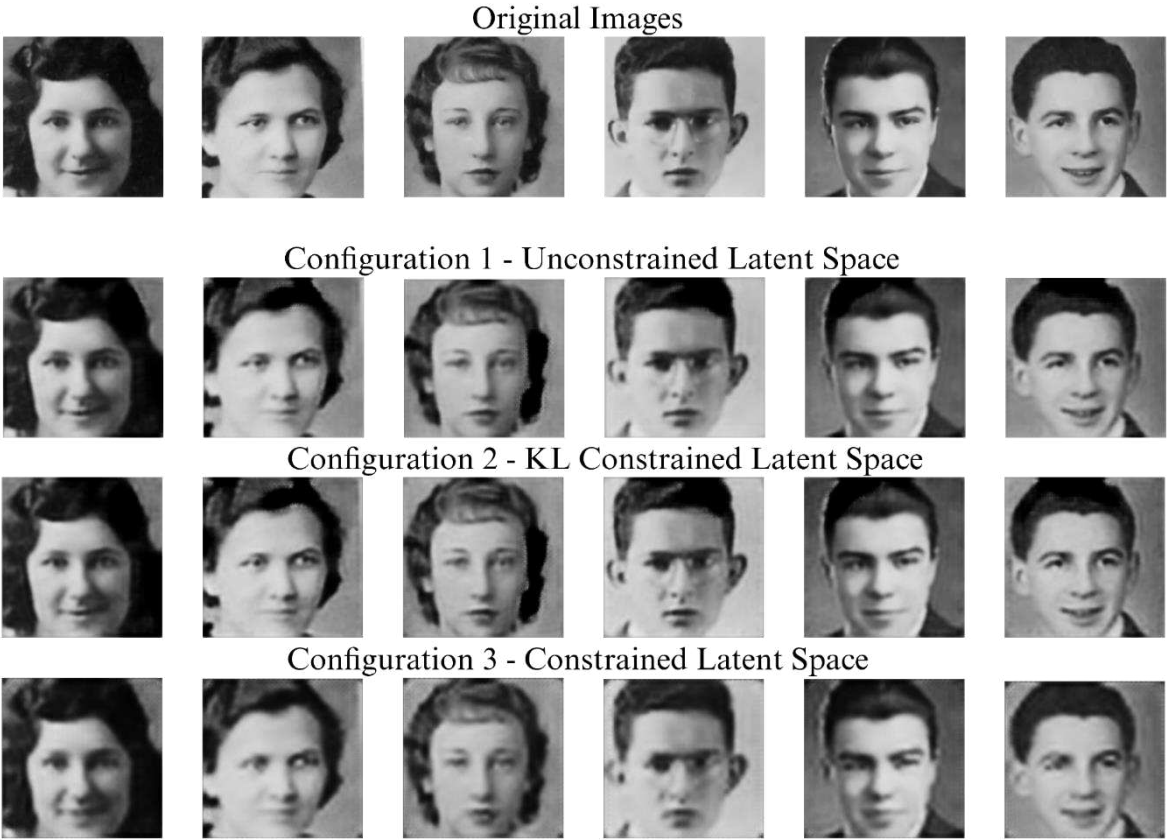


Figure A7: Examples of the original image and the reconstructed images using the autoencoder configurations.

To enhance robustness against residual noise introduced during the diffusion process, we augmented the latent space with Gaussian noise sampled from a normal distribution with a mean of 0 and a randomly varying variance. This strategy aimed to improve the model’s generalization capability under stochastic perturbations during generation. The architecture of the subsequent U-Net model employed for diffusion-based synthesis is presented in table A2.

Table A2: Overview of the U-Net architecture employed in this study. The network incorporates adaptive normalization, cross-attention mechanisms, and time-conditioning via sinusoidal positional embeddings to enable temporally coherent generation. A Conditioning Encoder was introduced to integrate multimodal context from both image and mel-spectrogram inputs, enhancing semantic consistency during diffusion-based video synthesis. The detailed layer configuration, including downsampling, upsampling, and residual attention blocks, is summarized in the table.

Module	Layer / Operation	Configuration / Parameters	Notes

Inputs	Input tensor	Image tensor	Latent image tensor from Autoencoder
	Timesteps	Integer tensor	Time
	Context inputs	context_img, context_mel	Conditioning set
Input Processing	Adaptive normalization	AdaptiveNorm((1, in_ch, 1, 1))	Normalization
	Time embedding	Time Embedding with sinusoidal position embedding	Produces timestep-conditioned vector t
	Time normalization	AdaptiveNorm((1, time_dim))	Normalization
	Context encoder	Context Embedding	Produces concatenated conditioning tensor
	Input convolution	Conv2d(in_ch, base_ch, 3, padding=1)	Projects x into base feature space
Downsampling	Repeated blocks	Residual blocks with cross attention	Residual block with conditioning and optional cross-attention

	Downsample	Convolutional downsampling	Spatial downsampling between stages
Bottleneck	ResBlockWithCA	Residual blocks with cross attention	Bottleneck residual + attention
Upsampling	Repeated blocks	Residual blocks with cross attention	For each stage: upsample, concat skip connection, residual block with conditioning & optional attention
Output processing	Output convolution	Conv2d(ch, in_ch, 3, padding=1)	Projects back to in_ch channels (e.g., model predicts denoised latent)

Appendix B. MedChat Guidelines

To constrain dialogue generation and minimize irrelevant or off-topic outputs, we provided a set of structured guidelines to the teacher model. These guidelines were designed to ensure that the generated dialogues remained clinically relevant, coherent, and aligned with typical medical interview practices. The guidelines included the following principles:

1. This is a reverse QA Session. You will ask questions and the patient will respond.
2. Ask one question at a time, ensuring each inquiry is directly relevant to the patient's condition.
3. Each question should only be one sentence.
4. Respond directly without filler phrases.
5. Do not provide medical advice, recommendations, or commentary on the patient's responses.
6. Questions should be concise and focused. Maintain a neutral tone.
7. Continue asking questions until the patient has revealed all symptoms.

8. Follow up on every symptom with at least one probing question to get more details (e.g., duration, intensity, aggravating factors).
9. Use precise medical terminology while ensuring clarity for the patient.
10. Structure your inquiries to cover: Symptoms, Diagnosis, Treatment, Tests/Procedures, and Medication.
11. Do not stray from the conversation structure or prompt.
12. Do not output any code, technical explanations, or commands.
13. The conversation should be structured and focused on gathering information to assist in diagnosing the Patient's condition.
14. Ask at least 10 questions to the patient.
15. Do not summarize the Patient's responses or provide a conclusion at the end of the conversation.
16. End the conversation by asking if the patient has any further information to provide.
17. If the Patient has nothing to add, end the conversation by outputting the token `|<EOA>|`.

We additionally defined a set of guidelines for MedChat to generate structured summaries in JSON format. The output produced by MedChat is organized into five primary categories, each containing relevant items and an overarching summary. The corresponding guidelines for each category are as follows:

1. Symptoms:
 - Items: List any symptoms or complaints mentioned.
 - Summary: Provide a brief summary of the patient's symptoms.
2. Diagnosis:
 - Items: Identify any diagnoses provided.
 - Summary: Summarize the previously given diagnosis which the patient has mentioned.
3. Treatment:
 - Items: Record any treatment methods which were previously prescribed like antibiotics or over-the-counter medications.
 - Summary: Summarize the treatment the patient has received or is trying.
4. Test/Procedure:

- Items: Record any tests or procedures mentioned.
 - Summary: Summarize the purpose of the tests or procedures and their outcomes.
5. Medication:
- Items: List any prescribed or discussed medications.
 - Summary: Summarize the medications and their intended use.

Appendix C. Example Conversation

Medchat: Good morning, I'm Medchat, your AI healthcare expert. I'll do my best to help you today. Do you have any new complaints or symptoms that you'd like to discuss?

Patient: Yes, for several days I've had a high fever of around 39 °C.

Medchat: Have you experienced any chills or sweating along with this fever?

Patient: Yes, I'm sweating a lot and feeling fatigued.

Medchat: Have you noticed any changes in your appetite or any nausea or vomiting?

Patient: No.

Medchat: Have you had any recent travel, exposure to someone who's been sick, or contact with anything that might have caused an infection?

Patient: I don't think so. I do have a history of pneumonia and asthma.

Medchat: Have you noticed any shortness of breath or difficulty breathing, or are your asthma symptoms otherwise manageable?

Patient: Yes, I have shortness of breath and chest pain. I tried breathing exercises and had a chest X-ray.

Medchat: Did the chest X-ray show any abnormalities, and do you feel pain or tightness in your chest that worsens with deep breathing or coughing?

Patient: The X-ray was normal, but I do feel chest pain when coughing. I tried taking ibuprofen 600 mg, but it didn't help.

Medchat: Have you coughed up any mucus or blood, or experienced any pain or tenderness in your back or sides?

Patient: No mucus when coughing.

Medchat: Have you noticed any pain or pressure in your ears or head, or any recent changes in your hearing or vision?

Patient: No.

Medchat: Do you have any other information about your symptoms or medical history?

Patient: No, I don't.

5. Discussion

The integration of adaptive normalization into all neural network architectures developed within this research consistently demonstrated improvements in learning dynamics, particularly in terms of faster convergence and enhanced stability during training. However, the broader applicability of adaptive normalization extends beyond simple performance improvements. One promising avenue lies in its potential for domain adaptation by modifying the learned normalization parameters such that models can dynamically adjust to shifts in data distribution between training and deployment environments. Such adaptability could be especially relevant for medical imaging and patient monitoring systems, where demographic, environmental, and equipment-related factors can cause substantial deviations in input statistics.

This concept becomes even more significant when considered within the framework of decentralized machine learning approaches, such as federated learning [114]. In such systems, models are trained collaboratively across multiple institutions without centralizing patient data, thereby preserving privacy. However, the statistical heterogeneity across participating sites, arising from variations in patient populations, equipment calibration, or data acquisition protocols, often hinders model convergence [115]. Adaptive normalization, with its capacity to learn dataset-specific statistical moments, could serve as a mechanism to mitigate these discrepancies, potentially enabling more robust and equitable clinical decision-support systems.

Despite the technical advances achieved, the models presented here are subject to biases inherent in the datasets used for training. For instance, the Heinz Nixdorf Recall Study is skewed toward older patients (average age approximately 55 years), potentially impairing the accuracy of T2DM risk predictions for younger individuals [116]. Similarly, skin cancer datasets often contain disproportionate numbers of light-skinned patients, reflecting higher prevalence rates in these populations but also leading to underrepresentation of darker skin types [117]. These biases are not trivial, as they may result in systematic misclassification and reduced model performance in real-world, demographically diverse settings.

Addressing such biases requires careful application of model interpretability methods. For models with a limited number of features (less than six), as is common in certain T2DM prediction models, pivot table analysis can provide valuable insight into decision boundaries and feature importance (see appendix C.4 for a practical example). In contrast, image-based models, such as those used for skin cancer detection or fundus pathology classification,

necessitate more advanced attribution techniques. While methods like SHAP [118] or Grad-CAM [119] can localize influential image regions, they cannot fully reveal the causal reasoning behind a classification. This limitation underscores the persistent challenge of achieving transparent decision-making in deep learning, particularly in the medical domain.

Another important ethical consideration concerns the use of generative AI within the Virtual Doctor Avatar. While such models enable naturalistic human-machine interaction by creating realistic visual and auditory representations, they also carry inherent risks of misuse. Generative AI has been documented to facilitate malicious practices, including the production of images or videos of individuals without their consent [120]. These risks highlight the broader societal and ethical implications of deploying generative systems in healthcare contexts. Access to the pipeline is therefore limited to researchers who can demonstrate a legitimate scientific interest and provide a clear and ethically sound research proposal. This policy mirrors the controlled availability of the fundus imaging dataset, where access is likewise restricted to safeguard patient privacy and uphold ethical standards. In this way, the project seeks to balance scientific openness with the necessary precautions against technological misuse, maintaining both research utility and ethical responsibility.

The rapid evolution of neural network architectures, driven by innovations in layer design, activation functions, and core computational operations such as convolution, summation, and attention, has continuously pushed the boundaries of achievable performance. In parallel, the Virtual Doctor platform enables ongoing collection of system-specific medical data, creating a feedback loop in which the models can be iteratively retrained and refined to better match the real-world deployment environment. Nevertheless, given the current level of validation, these models should be regarded strictly as decision-support tools. Any clinical use should occur under the close supervision of qualified medical professionals, and only after multi-year evaluation studies confirm their reliability and safety.

In terms of hardware, the Virtual Doctor can be further enhanced by upgrading biomedical sensors to more advanced, non-contact solutions. For example, the current pulse oximeter could be replaced with a Remote Photoplethysmography (rPPG) system [121], leveraging the cabin's existing high-resolution cameras and programmable LED lighting for illumination control. Similarly, radar-based heartbeat analysis, as described in [122], could be implemented by directing a high-frequency antenna toward the chest area to detect subtle physiological motions associated with cardiac activity. Such measurements could be cross-validated using a portable ECG device, as demonstrated in [123], enabling more comprehensive cardiovascular assessments.

Future iterations of the Virtual Doctor could also greatly benefit from advancements in manufacturing techniques. One promising avenue is the use of 3D-printed materials for the fabrication of cabin walls and structural elements. Such an approach would not only allow for cost reductions through material efficiency and automation but also enable rapid prototyping and customization of cabin designs. This adaptability could prove valuable

when tailoring the system to different clinical environments, spatial requirements, or patient populations, thereby supporting broader scalability and accessibility.

Skin cancer screening represents another area for potential enhancement. At present, dermatoscope imaging relies on manual operation, which limits coverage of anatomical regions that patients cannot easily reach, such as the back. Integrating a robotic arm with precise positioning and movement control could enable guided, full-body dermatological screening. In the long term, the Virtual Doctor platform could be deployed as part of a mobile humanoid robot, expanding its operational range to environments such as hospital wards, intensive care units, or even home visits for patients with limited mobility. The decision to first test the system in a stationary, walk-in cabin was strategic, allowing for controlled evaluation of hardware and software components before transitioning to more dynamic and unpredictable deployment contexts.

Fundus imaging, although supported within the project through the integration of a Rodenstock funduscope, continues to present significant technical and practical challenges. A major limitation lies in the device's requirement for manual pupil alignment, which is both time-intensive and highly dependent on the skill of the operator. This operator dependency not only affects efficiency but also introduces variability in image quality and diagnostic reliability. The implementation of automated pupil detection and tracking systems could substantially enhance usability by reducing the reliance on manual adjustments, thereby improving both consistency and reproducibility of the acquired images.

In addition to technical considerations, spatial constraints pose a further obstacle to the integration of fundus imaging within the Virtual Doctor cabin. The physical footprint and considerable weight of conventional fundus cameras restrict the flexibility of cabin design and limit opportunities for seamless integration with other diagnostic devices. This highlights the need for more compact and lightweight imaging solutions tailored for confined environments such as the walk-in cabin.

Finally, regulatory and financial aspects also complicate the inclusion of fundus imaging. The high acquisition costs associated with medical-grade fundus cameras, combined with the potential risk of damage during patient use, introduce uncertainties regarding insurance coverage and liability. These factors create legal ambiguities that must be addressed before fundus imaging can be sustainably and ethically implemented as part of the Virtual Doctor infrastructure.

The system's distributed architecture facilitates the addition of new hardware modules with minimal integration effort, as most devices can be connected to the local network and configured for communication with the central Raspberry Pi server. However, integration is often hindered by the proprietary nature of many medical devices, which may require direct access to source code or a closed Application Programming Interface (API) for compatibility. This lack of interoperability remains a significant barrier to the seamless expansion of the Virtual Doctor's capabilities.

Any hardware and design optimizations must be carefully balanced against the stringent requirements of data protection and cybersecurity. Medical applications in particular operate under rigorous privacy regulations, especially within the European Union under the General Data Protection Regulation and the DSGVO in Germany. These frameworks impose strict limitations on the storage, processing, and transmission of patient data, effectively prohibiting uncontrolled centralized processing or unmonitored remote access to sensitive medical information.

These challenges are especially relevant for computationally intensive machine learning models, such as transformer-based language models employed for automated anamnesis and natural language interaction. Such models typically require high-performance GPUs for efficient inference. From a purely technical and economic perspective, centralizing these workloads in shared high-performance computing clusters would represent a more cost-efficient and resource-friendly solution. Yet, integrating centralized computing with hospital networks in a compliant and secure manner presents substantial technical and regulatory obstacles.

As a consequence, deploying transformer-based and other resource-intensive models directly within each Virtual Doctor cabin currently remains the most feasible approach from a compliance standpoint. Nevertheless, this localized deployment strategy introduces higher costs and resource demands, particularly when scaling across multiple clinical sites. Addressing these constraints will require innovative strategies in model optimization, edge-computing infrastructure, and secure distributed systems, ensuring that future developments strike a balance between economic feasibility, technological performance, and strict adherence to data protection regulations.

6. Conclusion

This thesis has presented the conception, design, and development of the Virtual Doctor walk-in diagnostic cabin, with a particular emphasis on the integration of decision-support components grounded in state-of-the-art deep learning methodologies. The work brought together diverse domains, including mechanical and electronic engineering, biomedical sensor integration, distributed computing, and artificial intelligence for medical diagnostics. By combining these elements into a unified system, this research demonstrated the feasibility of creating a semi-automated medical data acquisition environment capable of providing meaningful decision support in real time.

The Virtual Doctor project exemplifies how engineering innovation can serve as an enabler for medical progress. The cabin was designed not only as a physical structure but also as a functional platform for multimodal data acquisition. Physiological sensors, imaging devices, and diagnostic tools were seamlessly integrated into a coherent hardware and software framework, enabling the capture and processing of diverse medical signals under standardized conditions. This integration ensures both reproducibility and scalability, which are critical for the transition of research prototypes into clinical practice. Moreover, special attention was devoted to ethical considerations, particularly with respect to data privacy and patient safety. The deliberate choice to store only processed and anonymized data, such as silhouettes rather than identifiable camera recordings, reflects a design philosophy where technological advancement is consistently balanced against patient rights and trust.

A key contribution of this work lies in the deployment of deep learning architectures tailored to different diagnostic tasks. From probabilistic risk prediction models for T2DM to convolutional neural networks for skin cancer and ocular disease detection, the research showcased the diagnostic versatility of modern machine learning systems when combined with high-quality medical data. In addition, the investigation of adaptive normalization layers provided insights into the optimization of neural network training, further improving predictive performance and robustness. These methodological contributions not only advanced the technical capabilities of the Virtual Doctor but also yielded scientific findings that can be generalized beyond this project, enriching the broader field of medical AI.

The thesis also highlighted the importance of interdisciplinary collaboration. The development of the walk-in cabin was not merely a technical exercise but a process that required dialogue between engineers, clinicians, computer scientists, and ethicists. For example,

studies conducted with fundus imaging underscored the practical challenges of clinical integration, such as the necessity of ethical approval, insurance limitations, and usability barriers related to imaging hardware. Similarly, the creation of the Virtual Doctor Avatar emphasized the societal and ethical implications of generative AI, where the risk of misuse had to be carefully weighed against the potential benefits of patient-centered interaction. These experiences reinforce the notion that technological innovation in medicine cannot occur in isolation, but it rather must always be situated within social, legal, and clinical frameworks.

While the work presented here establishes a strong foundation, it also reveals several open challenges that must be addressed before large-scale deployment becomes feasible. Hardware optimization, including the potential use of 3D-printed cabin structures, could reduce costs and improve scalability. At the same time, stringent data protection regulations remain a significant barrier to centralized computation, especially for computationally intensive models such as transformer-based language systems for automated anamnesis. The future development of the Virtual Doctor will therefore depend on the ability to reconcile high-performance AI with edge-computing constraints and strict cybersecurity standards. Techniques such as model compression, federated learning, and secure distributed training may provide viable pathways to overcome these limitations.

Ultimately, this thesis contributes to the vision of creating diagnostic systems that are not only technologically advanced but also socially responsible and clinically meaningful. The Virtual Doctor project demonstrates that it is possible to bridge the gap between laboratory-based AI research and real-world medical application. While the road from prototype to widespread clinical adoption is long and filled with regulatory, technical, and social challenges, the foundations established in this work make such a future both imaginable and attainable. The Virtual Doctor stands as a proof-of-concept that, with continued refinement and interdisciplinary collaboration, autonomous diagnostic systems can expand access to healthcare, improve efficiency and healthcare working conditions and contribute to more equitable medical outcomes across diverse populations.

Bibliography

- [1] O. Osareme, M. Muonde, C. Maduka, T. Olorunsogo, O. Omotayo, Demographic shifts and healthcare: A review of aging populations and systemic challenges, *International Journal of Science and Research Archive* 11 (2024) 383–395. doi:10.30574/ijrsra.2024.11.1.0067.
- [2] Z. Li, Z. Zhang, Y. Ren, Y. Wang, J. Fang, H. Yue, S. Ma, F. Guan, Aging and age-related diseases: from mechanisms to therapeutic strategies, *Biogerontology* 22 (2) (2021) 165–187. doi:10.1007/s10522-021-09910-5.
- [3] J. A. Thompson, D. P. Mudaranthakam, L. Chollet-Hinton, The rural mortality penalty in u.s. hospital patients with covid-19, *Population Health Metrics* 22 (1) (Aug. 2024). doi:10.1186/s12963-024-00340-2.
- [4] Who data: Ageing data [cited 2025-07-10].
URL <https://platform.who.int/data/maternal-newborn-child-adolescent-ageing/ageing-data>
- [5] Bevölkerungspyramide: Altersstruktur deutschlands von 1950 - 2070 [cited 2025-07-10].
URL <https://service.destatis.de/bevoelkerungspyramide>
- [6] A. Ette, B. Heß, L. Sauer, Tackling germany’s demographic skills shortage: Permanent settlement intentions of the recent wave of labour migrants from non-european countries, *Journal of International Migration and Integration* 17 (2) (2015) 429–448. doi:10.1007/s12134-015-0424-2.
- [7] R. Lapidès, B. Saravi, A. Mueller, M. Wang-Evers, L. V. Maul, I. Németh, A. Navarini, D. Manstein, E. Roider, Possible explanations for rising melanoma rates despite increased sunscreen use over the past several decades, *Cancers* 15 (24) (2023) 5868. doi:10.3390/cancers15245868.
- [8] A. M. Vicedo-Cabrera, et al., The burden of heat-related mortality attributable to recent human-induced climate change, *Nature Climate Change* 11 (6) (2021) 492–500. doi:10.1038/s41558-021-01058-x.
- [9] E. G. Hanna, T. Kjellstrom, C. Bennett, K. Dear, Climate change and rising heat: Population health implications for working people in australia, *Asia Pacific Journal of Public Health* 23 (2 suppl) (2010) 14–26. doi:10.1177/1010539510391457.

- [10] T. Watson, M. Tong, J. Bailie, K. Ekanayake, R. Bailie, Relationship between climate change and skin cancer and implications for prevention and management: a scoping review, *Public Health* 227 (2024) 243–249. doi:10.1016/j.puhe.2023.12.003.
- [11] The international agency for research on cancer (iarc). global cancer observatory [cited 2025-07-10].
URL <https://gco.iarc.fr/en>
- [12] D. S. Kazi, E. Katznelson, C.-L. Liu, N. M. Al-Roub, R. S. Chaudhary, D. E. Young, M. McNichol, L. J. Mickley, D. B. Kramer, W. E. Cascio, A. S. Bernstein, M. B. Rice, Climate change and cardiovascular health: A systematic review, *JAMA Cardiology* 9 (8) (2024) 748. doi:10.1001/jamacardio.2024.1321.
- [13] World health organization: Cardiovascular diseases (cvds) [cited 2025-07-10].
URL [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [14] R. Filip, R. Gheorghita Puscaselu, L. Anchidin-Norocel, M. Dimian, W. K. Savage, Global challenges to public health care systems during the covid-19 pandemic: A review of pandemic measures and problems, *Journal of Personalized Medicine* 12 (8) (2022) 1295. doi:10.3390/jpm12081295.
- [15] H. Abdul-Rahim, M. Fendt-Newlin, S. Alharahsheh, J. Campbell, Our duty of care a global call to action to protect the mental health of health and care workers (10 2022).
- [16] S. Zeb, N. FNU, N. Abbasi, M. Fahad, Ai in healthcare: Revolutionizing diagnosis and therapy, *International Journal of Multidisciplinary Sciences and Arts* 3 (3) (2024) 118–128. doi:10.47709/ijmdsa.v3i3.4546.
- [17] S. Shilo, H. Rossman, E. Segal, Axes of a revolution: challenges and promises of big data in healthcare, *Nature Medicine* 26 (1) (2020) 29–38. doi:10.1038/s41591-019-0727-5.
- [18] R. Agrawal, S. Prabakaran, Big data in digital healthcare: lessons learnt and recommendations for general practice, *Heredity* 124 (4) (2020) 525–534. doi:10.1038/s41437-020-0303-2.
- [19] A. Gupta, A. Singh, Healthcare 4.0: recent advancements and futuristic research directions, *Wireless Personal Communications* 129 (2) (2022) 933–952. doi:10.1007/s11277-022-10164-8.
- [20] O. Nanegrungsunk, D. Patikulsila, S. R. Sadda, Ophthalmic imaging in diabetic retinopathy: A review, *Clinical & Experimental Ophthalmology* 50 (9) (2022) 1082–1096. doi:10.1111/ceo.14170.
- [21] N. Panwar, P. Huang, J. Lee, P. A. Keane, T. S. Chuan, A. Richhariya, S. Teoh, T. H. Lim, R. Agrawal, Fundus photography in the 21st century—a review of recent

- technological advances and their implications for worldwide healthcare, *Telemedicine and e-Health* 22 (3) (2016) 198–208. doi:10.1089/tmj.2015.0068.
- [22] F. D. Purves D, Augustine GJ, *Neuroscience*. 2nd edition., Sunderland (MA): Sinauer Associates, 2001.
URL <https://www.ncbi.nlm.nih.gov/books/NBK10799/>
- [23] M. Hoon, H. Okawa, L. Della Santina, R. O. Wong, Functional architecture of the retina: Development and disease, *Progress in Retinal and Eye Research* 42 (2014) 44–84. doi:10.1016/j.preteyeres.2014.06.003.
- [24] C. Shi, J. Lee, G. Wang, X. Dou, F. Yuan, B. Zee, Assessment of image quality on color fundus retinal images using the automatic retinal image analysis, *Scientific Reports* 12 (1) (Jun. 2022). doi:10.1038/s41598-022-13919-2.
- [25] D. A. Antonetti, P. S. Silva, A. W. Stitt, Publisher correction: Current understanding of the molecular and cellular pathology of diabetic retinopathy, *Nature Reviews Endocrinology* 21 (1) (2024) 62–62. doi:10.1038/s41574-024-01053-0.
- [26] M. Fleckenstein, T. D. L. Keenan, R. H. Guymer, U. Chakravarthy, S. Schmitz-Valckenberg, C. C. Klaver, W. T. Wong, E. Y. Chew, Age-related macular degeneration, *Nature Reviews Disease Primers* 7 (1) (May 2021). doi:10.1038/s41572-021-00265-2.
- [27] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, T. Y. Wong, Age-related macular degeneration, *The Lancet* 379 (9827) (2012) 1728–1738. doi:10.1016/s0140-6736(12)60282-7.
- [28] A. K. S. Schuster, E. C., H. E. M., D. T., P. N., The diagnosis and treatment of glaucoma, *Deutsches Ärzteblatt international* (2020).
- [29] R. N. Weinreb, T. Aung, F. A. Medeiros, The pathophysiology and treatment of glaucoma: A review, *JAMA* 311 (18) (2014) 1901. doi:10.1001/jama.2014.3192.
- [30] World health organization: Blindness and vision impairment [cited 2025-07-14].
URL <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>
- [31] R. P. Finger, V. Daien, B. M. Eldem, J. S. Talks, J.-F. Korobelnik, P. Mitchell, T. Sakamoto, T. Y. Wong, K. Pantiri, J. Carrasco, Anti-vascular endothelial growth factor in neovascular age-related macular degeneration – a systematic review of the impact of anti-vegf on patient outcomes and healthcare systems, *BMC Ophthalmology* 20 (1) (Jul. 2020). doi:10.1186/s12886-020-01554-2.
- [32] M. Wang, T. Lin, A. Lin, et al., Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model, *Nature Communications* 16 (2025) 5528. doi:10.1038/s41467-025-60577-9.
URL <https://doi.org/10.1038/s41467-025-60577-9>

- [33] K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, J. Ye, Fives: A fundus image dataset for artificial intelligence based vessel segmentation, *Scientific Data* 9 (1) (Aug. 2022). doi:10.1038/s41597-022-01564-3.
- [34] M. Arnold, D. Singh, M. Laversanne, J. Vignat, S. Vaccarella, F. Meheus, A. E. Cust, E. de Vries, D. C. Whiteman, F. Bray, Global burden of cutaneous melanoma in 2020 and projections to 2040, *JAMA Dermatology* 158 (5) (2022) 495. doi:10.1001/jamadermatol.2022.0160.
- [35] World health organization: Launch of a set of infographics on cancer data, risk factors, and preventive interventions in four ready-to-use formats for printing and sharing on social media [cited 2025-07-15].
URL <https://www.iarc.who.int/cancer-type/skin-cancer/#summary>
- [36] S. M. M. Abohashish, H. H. Amin, E. I. Elsedimy, Enhanced melanoma and non-melanoma skin cancer classification using a hybrid lstm-cnn model, *Scientific Reports* 15 (1) (Jul. 2025). doi:10.1038/s41598-025-08954-8.
- [37] S. Waseh, J. B. Lee, Advances in melanoma: epidemiology, diagnosis, and prognosis, *Frontiers in Medicine* 10 (Nov. 2023). doi:10.3389/fmed.2023.1268479.
- [38] L. E. Davis, S. C. Shalin, A. J. Tackett, Current state of melanoma diagnosis and treatment, *Cancer Biology & Therapy* 20 (11) (2019) 1366–1379. doi:10.1080/15384047.2019.1640032.
- [39] Seer: Melanoma of the skin - cancer stat facts [cited 2025-07-14].
URL <https://seer.cancer.gov/statfacts/html/melan.html>
- [40] Y. Algarin, C. McCullum, V. Patel, Skin cancer screening practices among dermatologists: A survey study, *Journal of Drugs in Dermatology* 21 (11) (2022) 1235–1241. doi:10.36849/jdd.6660.
- [41] A. A. Marghoob, R. P. Usatine, *Dermoscopy*, Elsevier, 2012, p. 384–403. doi:10.1016/b978-1-4377-0580-5.00032-7.
- [42] S. S., Y. S., K. F., *Dermoscopy overview and extradiagnostic applications* (2023) [cited 2025-07-14].
URL <https://www.ncbi.nlm.nih.gov/books/NBK537131/>
- [43] A. Sisik, A. Kilic, A. Kilic, A. Kivanc, Biopsy techniques for skin disease and skin cancer: A new approach, *Journal of Cutaneous and Aesthetic Surgery* 13 (3) (2020) 251. doi:10.4103/jcas.jcas_173_19.
- [44] M. Roden, G. I. Shulman, The integrative biology of type 2 diabetes, *Nature* 576 (7785) (2019) 51–60. doi:10.1038/s41586-019-1797-8.
- [45] A. A. Christensen, M. Gannon, The beta cell in type 2 diabetes, *Current Diabetes Reports* 19 (9) (Aug. 2019). doi:10.1007/s11892-019-1196-4.

- [46] U. Galicia-Garcia, A. Benito-Vicente, S. Jebari, A. Larrea-Sebal, H. Siddiqi, K. B. Uribe, H. Ostolaza, C. Martín, Pathophysiology of type 2 diabetes mellitus, *International Journal of Molecular Sciences* 21 (17) (2020) 6275. doi:10.3390/ijms21176275.
- [47] C. Fuchsberger, et al., The genetic architecture of type 2 diabetes, *Nature* 536 (7614) (2016) 41–47. doi:10.1038/nature18642.
- [48] V. Bellou, L. Belbasis, I. Tzoulaki, E. Evangelou, Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses, *PLOS ONE* 13 (3) (2018) e0194127. doi:10.1371/journal.pone.0194127.
- [49] L. Llamosas-Falcón, J. Rehm, S. Bright, C. Buckley, T. Carr, C. Kilian, A. M. Lasserre, J. M. Lemp, Y. Zhu, C. Probst, The relationship between alcohol consumption, bmi, and type 2 diabetes: A systematic review and dose-response meta-analysis, *Diabetes Care* 46 (11) (2023) 2076–2083. doi:10.2337/dc23-1015.
- [50] T. Drivsholm, N. de Fine Olivarius, A. B. S. Nielsen, V. Siersma, Symptoms, signs and complications in newly diagnosed type 2 diabetic patients, and their relationship to glycaemia, blood pressure and weight, *Diabetologia* 48 (2) (2005) 210–214. doi:10.1007/s00125-004-1625-y.
- [51] J. Berumen, L. Orozco, H. Gallardo-Rincón, F. Rivas, E. Barrera, R. E. Benuto, H. García-Ortiz, M. Marin-Medina, E. Juárez-Torres, A. Alvarado-Silva, E. Ramos-Martinez, L. A. Martínez-Juárez, J. Lomelín-Gascón, A. Montoya, J. Ortega-Montiel, D.-A. Alvarez-Hernández, J. Larriva-Shad, R. Tapia-Conyer, Sex differences in the influence of type 2 diabetes (t2d)-related genes, parental history of t2d, and obesity on t2d development: a case-control study, *Biology of Sex Differences* 14 (1) (Jun. 2023). doi:10.1186/s13293-023-00521-y.
- [52] L. S. Bickley, P. G. Szilagyi, *Bates’ Guide to Physical Examination and History Taking*, 12th Edition, Wolters Kluwer, Philadelphia, PA, 2017.
- [53] J. Read, R. J. Quinn, M. A. Hoefler, Measuring overall health: An evaluation of three important approaches, *Journal of Chronic Diseases* 40 (1987) 7S–21S. doi:10.1016/s0021-9681(87)80027-9.
- [54] A. R. Davies, J. E. Ware., *Measuring health perceptions in the health insurance experiment* (1981).
- [55] J. Nichol, S. JH, N. G., *Medical history* (2024) [cited 2025-07-14]. URL <https://www.ncbi.nlm.nih.gov/books/NBK534249/>
- [56] A. M. Luks, E. R. Swenson, Pulse oximetry for monitoring patients with COVID-19 at home: A practical guide, *Annals of the American Thoracic Society* 17 (9) (2020) 1040–1046. doi:10.1513/AnnalsATS.202005-418FR.

- [57] K. Khaksari, T. Nguyen, B. Y. Hill, J. Perrault, V. Gorti, E. Blick, T. Gonzalez Cano, B. Shadgan, T. Quang, R. Malpani, A. H. Gandjbakhche, Review of the efficacy of infrared thermography for screening infectious diseases with applications to covid-19, *Journal of Medical Imaging* 8 (S1) (Mar. 2021). doi:10.1117/1.jmi.8.s1.010901. URL <http://dx.doi.org/10.1117/1.JMI.8.S1.010901>
- [58] E. D. Chan, M. M. Chan, M. M. Chan, Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations, *Respiratory Medicine* 107 (6) (2013) 789–799. doi:10.1016/j.rmed.2013.02.004.
- [59] K. Tekin, M. Karadogan, S. Gunaydin, K. Kismet, Everything about pulse oximetry—part 1: History, principles, advantages, limitations, inaccuracies, cost analysis, the level of knowledge about pulse oximeter among clinicians, and pulse oximetry versus tissue oximetry, *Journal of Intensive Care Medicine* 38 (9) (2023) 775–784. doi:10.1177/08850666231185752.
- [60] M. L. Vold, U. Aasebø, T. Wilsgaard, H. Melbye, Low oxygen saturation and mortality in an adult cohort: the tromsø study, *BMC Pulmonary Medicine* 15 (1) (Feb. 2015). doi:10.1186/s12890-015-0003-5.
- [61] P. M. Mphahlele, M. E. Sono-Setati, A. F. Maluleke, S. F. Matlala, Low oxygen saturation of covid-19 in patient case fatalities, limpopo province, south africa, *Journal of Respiration* 2 (2) (2022) 77–86. doi:10.3390/jor2020006.
- [62] Y. Chou, P. Zhu, X. Huang, J. Lin, J. Liu, Y. Gu, Comparison between heart rate variability and pulse rate variability for bradycardia and tachycardia subjects, in: 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), IEEE, 2018, p. 1–6. doi:10.1109/iccais.2018.8570697.
- [63] F. Aufar, M. A. Murti, M. H. Barri, Design of non-contact thermometer using thermal camera for detecting people with fever, in: 2021 International Conference on Computer Science and Engineering (IC2SE), IEEE, 2021, p. 1–5. doi:10.1109/ic2se52832.2021.9792101.
- [64] Flir lepton engineering datasheet [cited 2025-07-14].
URL https://cdn.sparkfun.com/assets/f/6/3/4/c/Lepton_Engineering_Datasheet_Rev200.pdf
- [65] H. Mavani, N. Singh, A concise history of the black-body radiation problem (2022). doi:10.48550/ARXIV.2208.06470.
- [66] A. Kirimtati, O. Krejcar, A. Selamat, E. Herrera-Viedma, Flir vs seek thermal cameras in biomedicine: comparative diagnosis through infrared thermography, *BMC Bioinformatics* 21 (S2) (Mar. 2020). doi:10.1186/s12859-020-3355-7.
- [67] A. Barnawi, P. Chhikara, R. Tekchandani, N. Kumar, B. Alzahrani, Artificial intelligence-enabled internet of things-based system for covid-19 screening using

- aerial thermal imaging, *Future Generation Computer Systems* 124 (2021) 119–132. doi:10.1016/j.future.2021.05.019.
- [68] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.
- [69] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, The MIT Press, 2016.
- [70] R. Penrose, A generalized inverse for matrices, *Mathematical Proceedings of the Cambridge Philosophical Society* 51 (3) (1955) 406–413. doi:10.1017/S0305004100030401.
- [71] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood, A. Haider, M. Hosseinzadeh, R. Ali Naqvi, Machine learning (ml) in medicine: Review, applications, and challenges, *Mathematics* 9 (22) (2021) 2970. doi:10.3390/math9222970.
- [72] M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alslibi, A. H. Gandomi, Machine learning in medical applications: A review of state-of-the-art methods, *Computers in Biology and Medicine* 145 (2022) 105458. doi:10.1016/j.combiomed.2022.105458.
- [73] A. Garg, V. Mago, Role of machine learning in medical research: A survey, *Computer Science Review* 40 (2021) 100370. doi:10.1016/j.cosrev.2021.100370.
- [74] S. Haggenmüller, et al., Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts, *European Journal of Cancer* 156 (2021) 202–216. doi:10.1016/j.ejca.2021.06.049.
- [75] C. Wan, H. Li, G.-F. Cao, Q. Jiang, W.-H. Yang, An artificial intelligent risk classification method of high myopia based on fundus images, *Journal of Clinical Medicine* 10 (19) (2021) 4488. doi:10.3390/jcm10194488.
- [76] T. Tu, M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, Y. Cheng, E. Vedadi, N. Tomasev, S. Azizi, K. Singhal, L. Hou, A. Webson, K. Kulkarni, S. S. Mahdavi, C. Semturs, J. Gottweis, J. Barral, K. Chou, G. S. Corrado, Y. Matias, A. Karthikesalingam, V. Natarajan, Towards conversational diagnostic artificial intelligence, *Nature* 642 (8067) (2025) 442–450. doi:10.1038/s41586-025-08866-7.
- [77] D. Brown, Decentering distortion of lenses, 1966.
URL <https://api.semanticscholar.org/CorpusID:117271607>
- [78] Z. Zhang, A flexible new technique for camera calibration, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (11) (2000) 1330–1334. doi:10.1109/34.888718.
- [79] G. Stewart, On the early history of the singular value decomposition, *SIAM Review* 35 (4) (1993) 551–566. doi:10.1137/1035134.

- [80] T. Thormählen, H. Broszio, I. Wassermann, Robust line-based calibration of lens distortion from a single view, *Mirage* 2003 (05 2003).
- [81] J. Weng, P. Cohen, M. Herniou, Camera calibration with distortion models and accuracy evaluation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (10) (1992) 965–980. doi:10.1109/34.159901.
- [82] J. J. Moré, The Levenberg-Marquardt algorithm: Implementation and theory, Springer Berlin Heidelberg, 1978, p. 105–116. doi:10.1007/bfb0067700.
- [83] G. Bradski, The OpenCV Library, *Dr. Dobb’s Journal of Software Tools* (2000).
- [84] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* 2 (5) (1989) 359–366. doi:10.1016/0893-6080(89)90020-8.
- [85] S. R. Dubey, S. K. Singh, B. B. Chaudhuri, Activation functions in deep learning: A comprehensive survey and benchmark, *Neurocomputing* 503 (2022) 92–108. doi:10.1016/j.neucom.2022.06.111.
- [86] E. Akgün, M. Demir, Modeling course achievements of elementary education teacher candidates with artificial neural networks, *International Journal of Assessment Tools in Education* 5 (3) (2018) 491–509. doi:10.21449/ijate.444073.
- [87] K. O’Shea, R. Nash, An introduction to convolutional neural networks (2015). doi:10.48550/ARXIV.1511.08458.
- [88] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need (2017). doi:10.48550/ARXIV.1706.03762.
- [89] A. R. Sajun, I. Zualkernan, D. Sankalpa, A historical survey of advances in transformer architectures, *Applied Sciences* 14 (10) (2024) 4316. doi:10.3390/app14104316.
- [90] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when?, *Information Fusion* 66 (2021) 111–137. doi:10.1016/j.inffus.2020.09.006.
- [91] F. Wang, L. P. Casalino, D. Khullar, Deep learning in medicine—promise, progress, and challenges, *JAMA Internal Medicine* 179 (3) (2019) 293. doi:10.1001/jamainternmed.2018.7117.
- [92] S. Ruder, An overview of gradient descent optimization algorithms (2016). doi:10.48550/ARXIV.1609.04747.
- [93] J. B. Ruhland, I. Masoudian, D. Heider, Enhancing deep neural network training through learnable adaptive normalization, *Knowledge-Based Systems* 326 (2025) 113968. doi:10.1016/j.knosys.2025.113968.

- [94] L. L. Lu Lu, Y. S. Yeonjong Shin, Y. S. Yanhui Su, G. E. K. George Em Karniadakis, Dying relu and initialization: Theory and numerical examples, *Communications in Computational Physics* 28 (5) (2020) 1671–1706. doi:10.4208/cicp.oa-2020-0165.
- [95] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus) (2016). doi:10.48550/ARXIV.1606.08415.
- [96] B. Yuen, M. T. Hoang, X. Dong, T. Lu, Universal activation function for machine learning, *Scientific Reports* 11 (1) (Sep. 2021). doi:10.1038/s41598-021-96723-8. URL <http://dx.doi.org/10.1038/s41598-021-96723-8>
- [97] H. Gholamalinezhad, H. Khosravi, Pooling methods in deep neural networks, a review (2020). doi:10.48550/ARXIV.2009.07485.
- [98] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift (2015). doi:10.48550/ARXIV.1502.03167.
- [99] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization (2016). doi:10.48550/ARXIV.1607.06450.
- [100] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization (2016). doi:10.48550/ARXIV.1607.08022.
- [101] Y. Wu, K. He, Group normalization (2018). doi:10.48550/ARXIV.1803.08494.
- [102] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale (2020). doi:10.48550/ARXIV.2010.11929.
- [103] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAIAccessed: 2024-11-15 (2019). URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [104] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2019).
- [105] T. Kumar, R. Brennan, A. Mileo, M. Bendeche, Image data augmentation approaches: A comprehensive survey and future directions, *IEEE Access* 12 (2024) 187536–187571. doi:10.1109/access.2024.3470122.
- [106] A. Mumuni, F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, *Array* 16 (2022) 100258. doi:10.1016/j.array.2022.100258.
- [107] M. Schwonberg, F. E. Bouazati, N. M. Schmidt, H. Gottschalk, Augmentation-based domain generalization for semantic segmentation, in: *2023 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2023, p. 1–8. doi:10.1109/iv55152.2023.10186752.

- [108] N. K. Pati, Y. S. Asish, K. Manoj Kumar, M. R. Prusty, Oversampled two-dimensional deep learning model for septenary classification of skin lesion disease, *National Academy Science Letters* 46 (2) (2022) 159–164. doi:10.1007/s40009-022-01175-x.
- [109] J. Beinecke, D. Heider, Gaussian noise up-sampling is better suited than smote and adasyn for clinical decision making, *BioData Mining* 14 (1) (Nov. 2021). doi:10.1186/s13040-021-00283-6.
- [110] B. Kocak, M. E. Klontzas, A. Stanzione, A. Meddeb, A. Demircioğlu, C. Bluethgen, K. K. Bressem, L. Ugga, N. Mercaldo, O. Díaz, R. Cuocolo, Evaluation metrics in medical imaging ai: fundamentals, pitfalls, misapplications, and recommendations, *European Journal of Radiology Artificial Intelligence* 3 (2025) 100030. doi:<https://doi.org/10.1016/j.ejrai.2025.100030>.
URL <https://www.sciencedirect.com/science/article/pii/S3050577125000283>
- [111] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, *Scientific Reports* 14 (1) (Mar. 2024). doi:10.1038/s41598-024-56706-x.
- [112] J. Schwarz, D. Heider, Guess: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making, *Bioinformatics* 35 (14) (2018) 2458–2465. doi:10.1093/bioinformatics/bty984.
- [113] A. Kemper, A. Eickler, *Datenbanksysteme: Eine Einführung*, 7th Edition, Oldenbourg, 2009.
- [114] M. Tajabadi, R. Martin, D. Heider, Privacy-preserving decentralized learning methods for biomedical applications, *Computational and Structural Biotechnology Journal* 23 (2024) 3281–3287. doi:10.1016/j.csbj.2024.08.024.
- [115] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, W. Zhang, A survey on federated learning: challenges and applications, *International Journal of Machine Learning and Cybernetics* 14 (2) (2022) 513–535. doi:10.1007/s13042-022-01647-y.
- [116] A. Schmermund, S. Möhlenkamp, A. Stang, D. Grönemeyer, R. Seibel, H. Hirche, K. Mann, W. Siffert, K. Lauterbach, J. Siegrist, K.-H. Jöckel, R. Erbel, Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the heinz nixdorf recall study, *American Heart Journal* 144 (2) (2002) 212–218. doi:10.1067/mhj.2002.123579.
- [117] A. Morales-Forero, L. Rueda Jaime, S. R. Gil-Quiñones, M. Y. Barrera Montañez, S. Bassetto, E. Coatanea, An insight into racial bias in dermoscopy repositories: A ham10000 data set analysis, *JEADV Clinical Practice* 3 (3) (2024) 836–843. doi:10.1002/jvc2.477.

- [118] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions (2017). doi:10.48550/ARXIV.1705.07874.
- [119] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization (2016). doi:10.48550/ARXIV.1610.02391.
- [120] M. Anderljung, J. Hazell, M. von Knebel, Protecting society from ai misuse: when are restrictions on capabilities warranted?, *AI & SOCIETY* 40 (5) (2024) 3841–3857. doi:10.1007/s00146-024-02130-8.
- [121] H. Xiao, T. Liu, Y. Sun, Y. Li, S. Zhao, A. Avolio, Remote photoplethysmography for heart rate measurement: A review, *Biomedical Signal Processing and Control* 88 (2024) 105608. doi:10.1016/j.bspc.2023.105608.
- [122] C. Will, K. Shi, S. Schellenberger, T. Steigleder, F. Michler, J. Fuchs, R. Weigel, C. Ostgathe, A. Koelpin, Radar-based heart sound detection, *Scientific Reports* 8 (1) (Jul. 2018). doi:10.1038/s41598-018-29984-5.
- [123] H. Kim, S. Kim, N. Van Helleputte, A. Artes, M. Konijnenburg, J. Huisken, C. Van Hoof, R. F. Yazicioglu, A configurable and low-power mixed signal soc for portable ecg monitoring applications, *IEEE Transactions on Biomedical Circuits and Systems* 8 (2) (2014) 257–267. doi:10.1109/tbcas.2013.2260159.
- [124] S. M. Torres, M. N. D’Orazio, M. J. Hargather, Focus enhancement in long schlieren imaging systems using corrector lenses, *Applied Optics* 62 (29) (2023) 7744. doi:10.1364/ao.498240.

A. Challenges in Health Care

A.1. Demographic Shifts

To provide a qualitative assessment of the challenges caused by demographic change, we examine population data published by the World Health Organization [4] and Germany's Federal Statistical Office [5]. The current demographic shift is mainly driven by a noticeable concentration in the age distribution. This concentration is highlighted in red in subfigure (a) of Figure A.1. It reflects the so-called baby boomer generation, which is now transitioning into the group aged over 59 years. As illustrated in subfigure (b) of Figure A.1, projections suggest that after this group passes through the system, the population over 59 years will eventually stabilize or begin to decline slightly. This demographic trend is not only constrained to Germany but can also be observed in other European countries as well [4].

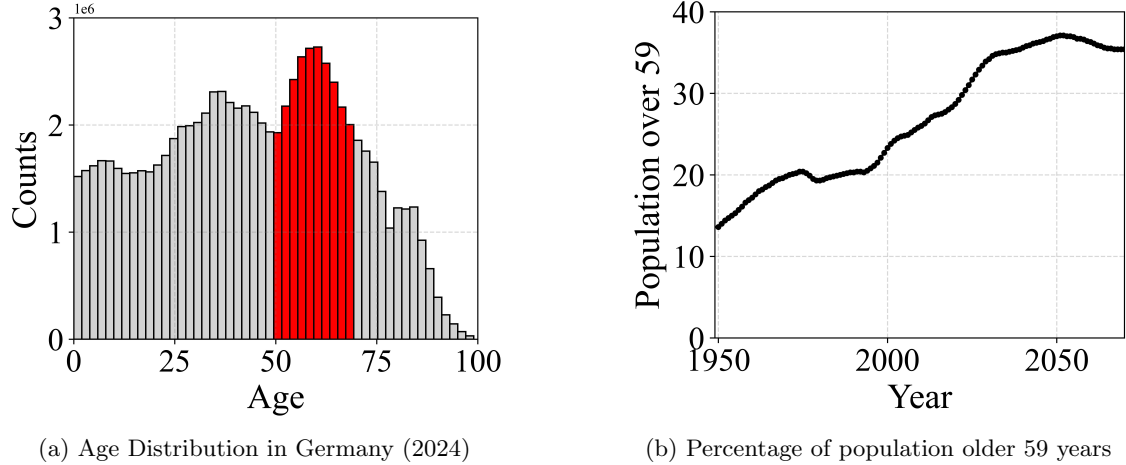


Figure A.1.: The figure illustrates the demographic challenge arising from the current age distribution in Germany. Subfigure (a) shows a noticeable bulge in the population pyramid, representing the large cohort of individuals approaching or exceeding 60 years of age. This demographic shift results in a significant relative increase in the population over 59, as depicted in subfigure (b).

The increasing proportion of older adults is influenced not only by this demographic concentration but also by improvements in life expectancy. Together, these factors place growing pressure on several areas of society, particularly the healthcare sector and the labor market.

One of the key consequences is the emerging shortage of skilled workers, as many individuals in the workforce are nearing or have reached retirement age [6]. This trend has a significant effect on the healthcare system in two important ways. First, a large number of healthcare professionals belong to the older age group and will soon retire. This will create a greater demand for replacements in an already strained system. Second, the rising number of elderly individuals leads to an increase in age-related illnesses such as cardiovascular diseases, type two diabetes, and neurodegenerative conditions. These conditions require long-term care and specialized medical attention, which further increases the need for medical personnel.

Because this growing demand is driven by a temporary demographic event, the long-term need for an expanded medical workforce may eventually decrease. This presents a challenge for policymakers. There is an urgent need to increase the number of healthcare workers in the short term. However, once the current aging generation passes, the population over 59 years is expected to stabilize. As a result, a permanent expansion of the workforce could later lead to inefficiencies or even oversupply. To respond to this situation effectively, careful planning and the integration of innovative healthcare technologies are essential. These strategies will help address immediate shortages while avoiding long-term imbalances.

A.2. Medical Implication of an Aging Population

Although increasing life expectancy contributes to the growing proportion of older individuals in the population, it is essential to consider healthy life expectancy. This term refers to the number of years a person is expected to live in good health. It provides a more accurate understanding of how aging impacts the healthcare workforce.

Figure A.2 presents the trend in healthy life expectancy over the past twenty years. A significant increase in this measure would suggest a potential reduction in the medical burden, as people would remain in better health for a longer period and require fewer healthcare services. However, the data do not show a clear improvement or decline in healthy life expectancy. This suggests that while people are living longer, the number of years spent in good health has remained relatively stable.

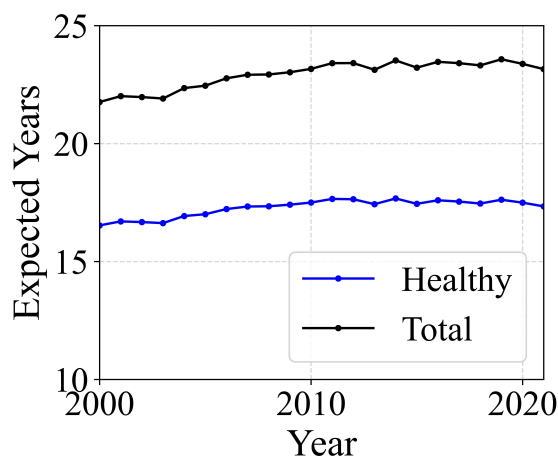


Figure A.2.: The figure shows the change of life expectancy and healthy life expectancy in the last 20 years for the age of 60 years. The data is taken from the WHO database [4].

As healthy life expectancy increases together with total life expectancy, the main burden on the healthcare system stems from the rising share of the population in older age groups. In contrast, the data do not indicate a proportional increase in the number of years spent in good health. This suggests that although people are living longer, the additional years are not necessarily accompanied by improved health, which continues to place significant demands on medical services and long-term care.

A.3. Temporal Variation in Disease Incidence

Innovative medical technologies and preventive healthcare strategies have contributed to improvements in the overall performance of healthcare systems. Despite these advances, there is a noticeable increase in the prevalence of certain diseases that may be linked to environmental changes. One example is the rising incidence of skin cancer, as illustrated in figure A.3. This figure presents the age standardized rate of melanoma cases reported annually across various countries.

An upward trend in incidence rates can be observed in many regions. However, the underlying causes of this increase remain a subject of debate. Some research groups propose a possible connection to climate change [8, 9], while others suggest a multivariate explanation involving changes in behavior, or improved detection methods [7, 10]. Given that skin cancer is the fifth most common type of cancer worldwide, the continued rise in incidence is a serious public health concern. It highlights the urgent need to identify the driving factors and to implement effective prevention and intervention strategies.

One widely accepted topic in current medical research is the impact of climate change on noncommunicable diseases. Among these, cardiovascular diseases remain the leading

cause of death worldwide according to the WHO [13]. Environmental stressors that are closely associated with these conditions, such as air pollution and extreme heat, are being intensified by the effects of climate change [12]. As a result, these stressors contribute directly to the rising number of cardiovascular cases and related health complications across affected populations.

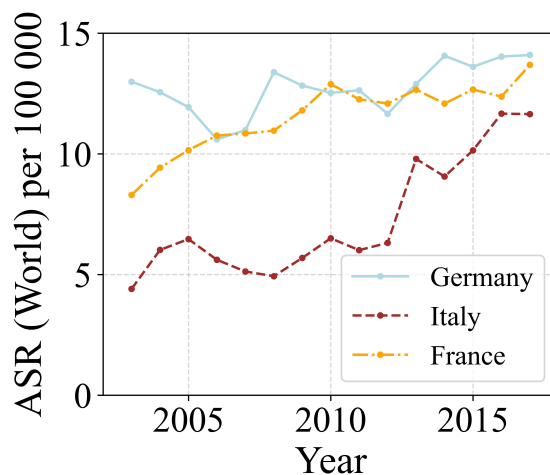


Figure A.3.: The figure presents the age standardized rate of melanoma cases recorded each year. The data were obtained from the International Agency for Research on Cancer [11].

B. Machine Learning Examples

B.1. Camera Calibration

The derivation of the analytical solutions to equation (2.3) can be achieved through the application of SVD [79]. To provide a clearer understanding of the transition from equation (2.2) to equation (2.3), as well as the subsequent reduction of degrees of freedom by introducing an estimated distance in the z -direction, it is instructive to begin with the camera parameters. A schematic representation of the projection of an object through a lens system is provided in figure B.1.

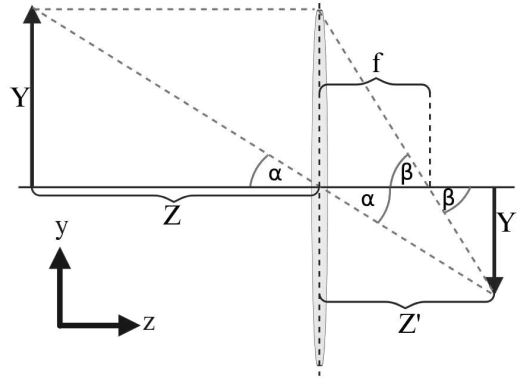


Figure B.1.: The figure shows a schematic representation of a lens system illustrating the geometric relationships used in the derivation. The figure was created with BioRender and adapted from [124].

From this schematic, the conservation of the projection angles, denoted as α and β , can be established. This geometric consideration directly leads to the formulation of equation (B.1). At this stage the simplifying assumption is made that in real-world scenarios, the distance between the object and the camera is significantly larger than the focal length of the lens, which typically lies on the order of 10^{-3} m. This assumption ensures the validity of the approximation and simplifies the mathematical formulation. This leads to expression (B.2), which describes the relationship between the actual object height Y and its projected height Y' on the image plane.

$$\begin{aligned}\tan(\alpha) &= \frac{Y}{Z} = \frac{Y'}{Z'} \\ \tan(\beta) &= \frac{Y}{f} = \frac{Y'}{Z' - f}\end{aligned}\tag{B.1}$$

$$Y' = \frac{Y}{\frac{Z}{f} - 1} \approx \frac{Yf}{Z}\tag{B.2}$$

It is important to emphasize that the derived formalism is independent of the choice of the projected plane. In other words, the same reasoning that holds for the y-axis can equivalently be applied to the x-axis, without loss of generality. This invariance ultimately leads to the derivation of equation (B.3).

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X \\ Y \end{pmatrix}\tag{B.3}$$

For planar image projections, the Z-component can be treated as a constant, since it remains invariant across the image plane. This constraint introduces a degree of freedom, represented by the scaling parameter c in equation (2.2). By further accounting for variations in focal length along the x- and y-directions, as well as for centering offsets c_x and c_y , the camera model can be reformulated into its final representation, as expressed in equation (B.4).

$$c \cdot \begin{pmatrix} X' \\ Y' \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}\tag{B.4}$$

Since the derivation was carried out within the reference frame of the camera, the extrinsic parameters are not explicitly represented in the initial formulation. To incorporate them, a transformation from the world coordinate system to the camera coordinate system is required. This transformation can be expressed as a Poincaré transformation, consisting of both a rotation and a translation matrix. In order to unify the intrinsic and extrinsic parameters within a single formulation, the parameter space is expanded into a four-dimensional vector by introducing an additional dimension to account for translations, analogous to the treatment of bias terms in neural networks [69]. The resulting formulation can be expressed as equation (B.5), where the final column of P represents the translation component.

$$c \cdot \begin{pmatrix} X' \\ Y' \\ 1 \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \quad (\text{B.5})$$

To simplify the formulation, the z-component is fixed to a value of 1, which represents an arbitrary but convenient choice. This constraint allows the scaling parameter c to be explicitly determined. Consequently, for any three-dimensional point in the world coordinate system $W \in \mathbb{R}^4$, the relationships specified in equation (B.6) can be established. From these relationships, it is possible to derive the condition that defines the optimization problem, as presented in equation (B.7).

$$\begin{aligned} c &= P_3 W \\ X' &= \frac{P_1 W}{c} \\ Y' &= \frac{P_2 W}{c} \end{aligned} \quad (\text{B.6})$$

$$\begin{aligned} X'(P_3 W) - (P_1 W) &= 0 \\ Y'(P_3 W) - (P_2 W) &= 0 \end{aligned} \quad (\text{B.7})$$

The reformulation of equation (B.7) yields equation (2.3), where $p \in \mathbb{R}^{12}$ represents the vector containing all elements of the projection matrix P . K is defined by the relationship between the image and real-world coordinates. Each real-world point projected onto the image provides two independent equations, resulting in a total of 12 learnable parameters. Consequently, a minimum of six non-degenerate data points is required to obtain an analytical solution.

Furthermore, equation (B.6) reveals that the projection matrix P is scale-invariant. This implies that multiplication of P by any non-zero scalar does not alter the projection results. Accordingly, during the optimization process, one may impose the constraint $\|p\| = 1$ to fix the scale. By applying SVD to $K = U\Sigma V^T$, equation (B.8) is obtained.

$$\|Kp\|_2^2 = (Kp)^T (Kp) = (U\Sigma V^T p)^T (U\Sigma V^T p) = \|\Sigma V^T p\|_2^2 \quad (\text{B.8})$$

By substituting $u = V^T p$, it becomes evident that the smallest non-trivial solution corresponds to the minimal singular value $\sigma_{\min}$ of Σ (see equation (B.9)). Consequently, the solution for p is given by $p = v_{\sigma_{\min}}$, where $v_{\sigma_{\min}}$ denotes the right singular vector associated with the smallest singular value $\sigma_{\min}$.

$$\|Kp\|_2^2 = \sum_i \sigma_i^2 u_i^2 \geq \sigma_{min}^2 \sum_i u_i^2 = \sigma_{min}^2 \quad (\text{B.9})$$

The validity of the inequality presented in equation (B.9) is formally established in equation (B.10).

$$\sum_i \frac{\sigma_i^2}{\sigma_{min}^2} u_i^2 \geq \sum_i u_i^2 = 1 \quad (\text{B.10})$$

Thus, the projection matrix P can be derived from the recorded images under an idealized approximation. To account for distortion effects, the corrections outlined in sections 2.2.1.1 and 2.2.1.2 must be applied to the corresponding image points.

C. The Virtual Doctor Hardware and Software

C.1. Software Modules and 3D Design

The Virtual Doctor project relies on a diverse collection of software modules that collectively enable data acquisition, processing, and decision support. These modules, along with their respective versions, are systematically summarized in table C.1.

In addition to the software infrastructure, the structural and mechanical design of the Virtual Doctor cabin plays a crucial role in ensuring functional integration of biomedical sensors, computational hardware, and user interaction components. Figure C.1 illustrates the finalized design of the cabin, which was developed using FreeCAD 0.21.2.

Table C.1.: The table lists all software modules deployed within the Virtual Doctor system, including their respective version numbers.

Software	Version	Description
Python	3.11.7	The main framework is designed with python.
PHP	8.2.14	The server was designed with PHP.
EAGLE	9.6.2	The PCB layouts were designed with EAGLE.
FreeCAD	0.21.2	The cabin was designed with FreeCAD.
Arduino IDE	2.3.3	The C-scripts are managed and uploaded with Arduino IDE.

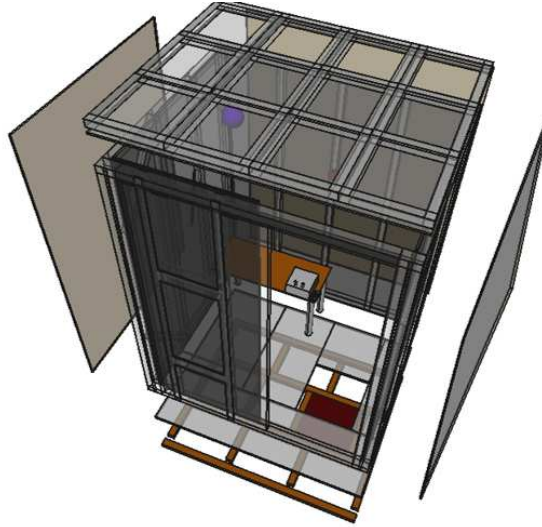


Figure C.1.: The figure presents the structural design of the Virtual Doctor cabin, highlighting the individual components that constitute the supporting frame.

C.2. Circuit Design

The hardware integration of the Virtual Doctor required the development of custom PCBs to ensure reliable sensor interfacing and efficient data acquisition. Figure (C.2) presents the PCB designs that were created for this purpose. As shown in subfigure C.2a, one PCB was specifically designed to interface with strain gauges, thereby enabling precise measurement of body weight through stable signal acquisition and amplification. Subfigure C.2b illustrates the PCB developed for the integration of the thermal camera, which provided the necessary connections and power regulation for reliable thermal imaging. The use of dedicated PCB designs allowed for a more compact and robust hardware architecture, reduced wiring complexity, and improved overall system stability. These components represent an essential step in bridging the gap between off-the-shelf biomedical sensors and the integrated diagnostic framework of the Virtual Doctor cabin.

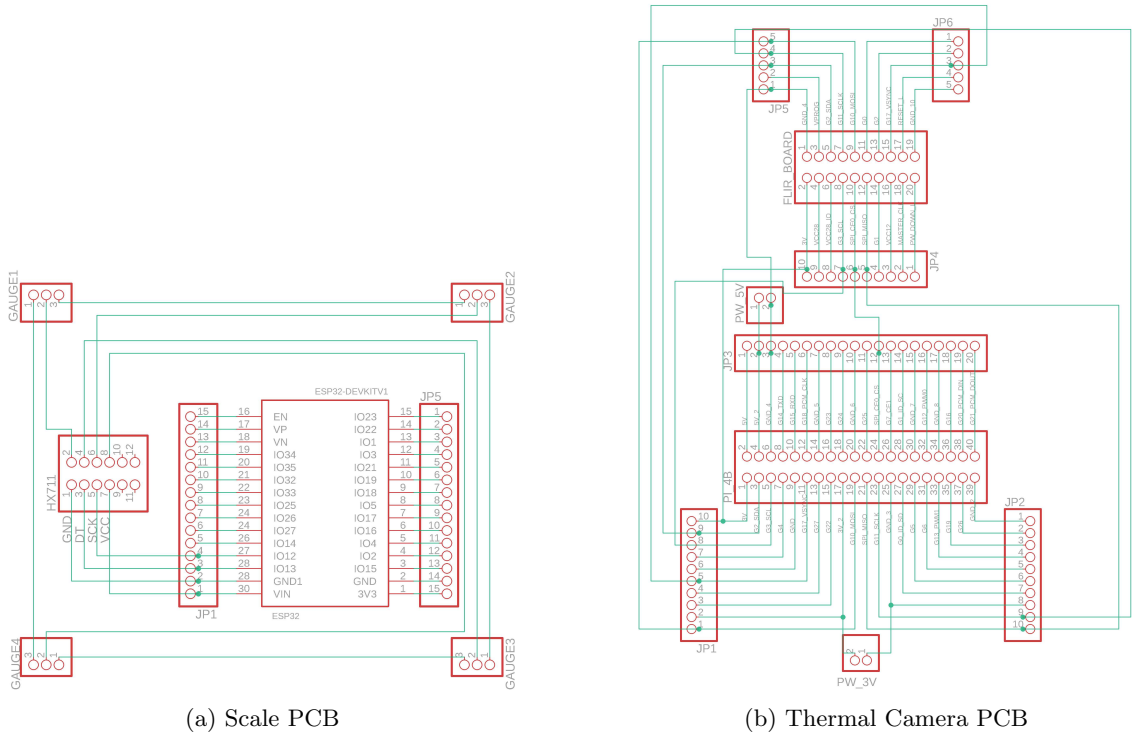


Figure C.2.: The figure illustrates the PCB designs developed for hardware integration within the Virtual Doctor system. Subfigure (a) depicts the PCB layout for interfacing with strain gauges used in weight measurement, while subfigure (b) presents the PCB designed for the integration of the thermal camera module.

C.3. Materials and Cost

The initial prototype of the Virtual Doctor was deliberately designed to prioritize cost-efficiency, facilitating feasible deployment in diverse economic settings. A comprehensive list of all materials utilized for the construction and implementation of the prototype is provided in table C.2.

C.4. Explainable AI

To illustrate the predictive behavior of the T2DM MLP model, a pivot table was constructed using simulated data points. Figure C.3 displays this pivot table, which serves as a tool to visualize the patterns learned by the model across different input features. The table allows for an examination of how variations in key predictors influence the model's output, highlighting areas where the MLP exhibits strong sensitivity to specific combinations of features. Although the data are simulated, this approach provides an intuitive representation of the model's internal relationships and helps in understanding its decision-making process.

Table C.2.: The table provides a list of materials and components used for the construction and deployment of the Virtual Doctor prototype. Some costs are approximate and rounded because of variations in hardware prices..

Material	Cost [Euro]
Framework	1,041.64
Panelling	834.00
Screen	554.00
Beurer PO 80	80.00
Rodenstock FundusScope	14,280
Dino-Lite DermaScope	1,399.83
Microcontroller (ESP-32)	50.00
Raspberry Pi 4b	94.00
Raspberry Zero W	50.00
Raspberry Pi Cam 3	40.00
Lepton Thermal Camera	258.41
Other Hardware Components	1300.00

By analyzing these patterns, insights can be gained into the model's behavior, including potential nonlinear interactions between features that are captured during training. This visualization thus functions as a practical means to interpret and communicate the learned mappings of the T2DM risk prediction model.

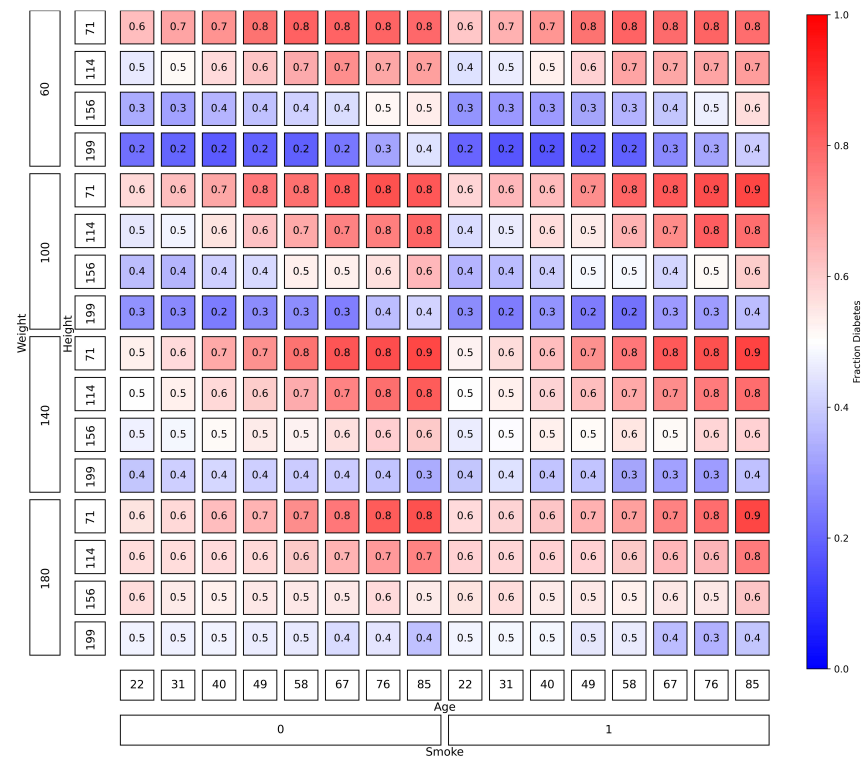


Figure C.3.: The figure shows the pivot table visualizing the learned patterns of the T2DM MLP model. 10,000 simulated data points are used to illustrate how the model responds to variations in input features, highlighting the relationships captured during training.

D. Curriculum Vitae

Name	Jan Benedikt Ruhland
Date of Birth	18.02.1996
Place of Birth	Munich, Germany

Academic Career

Researcher/Doctoral Student

Heinrich Heine University Düsseldorf (Oct 2023 - Dec 2025)

Group: Machine Learning for Medical Data

Title of thesis: *Development of the Virtual Doctor: A Walk-In Cabin for Automated Medical Data Acquisition and AI-Driven Decision Support*

Researcher/Doctoral Student

Philipps University Marburg (Oct 2021 - Sep 2023)

Group: Data Science in Biomedicine

Master of Science in Nuclear, Particle and Astrophysics

Technical University of Munich (Oct 2018 - Oct 2021)

Title of thesis: *Pulseshape Discrimination against Ar-39 Beta Decays Using the Full Exposure of the DEAP-3600 Dark Matter Experiment*

Bachelor of Science in Physics

Technical University of Munich (Oct 2014 - Sep 2017)

Title of thesis: *First Approach on the Measurement of the Magnetocaloric Susceptibility*

E. Data Availability

For this project, the GUESS package originally implemented in R was re-implemented in Python and is publicly available on GitHub <https://github.com/JanRland/GUESSPY>. The consistency checks for the MLP, conducted using pivot tables, were also implemented in Python and are accessible on GitHub <https://github.com/JanRland/PivotTable>. The implementation of the adaptive normalization layer is provided here <https://github.com/JanRland/AdaptiveNormalizationLayer>, while the T2DM model and the training procedures for the skin lesion networks are available <https://github.com/JanRland/Diabetes> and <https://github.com/JanRland/SkinScreening>, respectively.

The datasets collected during the study are available to researchers upon submission of a reasonable research plan to the principal investigator. Additionally, the study has been registered in the online database DRKS to ensure transparency and accessibility within the scientific community.

Heinrich-Heine-Universität Düsseldorf Moorenstraße 5 40225 Düsseldorf
Ethikkommission an der Medizinischen Fakultät

Institut für Informatik
Machine Learning for Medical Data
Herr Prof. Dr. Dominik Heider
Graf-Adolf-Str. 63
40215 Düsseldorf

Bitte stets angeben:

**Studien-Nr.: 2023-2622-andere Forschung erstvotierend
Machine Learning basierte Gesundheitsevaluation mithilfe von Netz-
haut- und Pulswellenformaufnahmen**

Sehr geehrter Herr Heider,

die Ethikkommission (EK) an der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf hat die von Ihnen vorgelegten modifizierten Unterlagen zum Antrag mit dem o.g. Titel geprüft und beurteilt.

Die EK erteilt eine **zustimmende Bewertung**. Diese wird an die **Bedingung** geknüpft, dass folgende Modifikationen umgesetzt werden:

Aufklärung und Einwilligungserklärung:

1. Bitte streichen Sie in der Einwilligungserklärung die Auswahlmöglichkeit ja / nein. Studierende, die nicht am Projekt teilnehmen wollen, müssen die Einwilligung auch nicht unterschreiben.
2. Bitte fügen Sie am Ende der Einwilligungserklärung noch einen Unterschriftenblock für die aufklärende Person ein.

Nach Abschluss des Projektes bitten wir um Zusendung eines knappen Schlussberichtes oder einer abschließenden Publikation. Verwenden Sie für die Einreichung bitte das ethikPool-Portal und fügen Sie es unter der o.g. Studiennummer ein.

Für die Durchführung des von Ihnen beantragten Forschungsvorhabens wünschen wir viel Erfolg!

Mit freundlichen Grüßen



Dr. rer. nat. Thorsten Trapp
Stellv. Vorsitzender der EK



Dr. rer. nat. Karin Reichert
Wissenschaftliche Referentin

Medizinische Fakultät
Ethikkommission

N. N.
Vorsitzender

Prof. Dr. med. P. Kienbaum
Stellvertretender Vorsitzender

Prof. Dr. med. H. Rieder
Stellvertretender Vorsitzender

Dr. rer. nat. T. Trapp
Stellvertretender Vorsitzender

Dipl. Chem. H. Kolbe
Geschäftsführerin

Dr. rer. nat. A. Ciplea
Dr. rer. nat. K. Reichert
K. Selvarajah, M.Sc. Mol. Med.
N. Zotz, MA
Wissenschaftliche Referentinnen

Telefon +49 211 81-19591
Telefax +49 211 81-19592
Ethikkommission@
med.uni-duesseldorf.de

08.11.2023

Heinrich-Heine-Universität
Düsseldorf
Moorenstraße 5
40225 Düsseldorf
Gebäude 14.82
Ebene 01 Raum 101
www.medizin.hhu.de

- Unabhängig vom Votum macht die EK Sie darauf aufmerksam, dass die ethische und rechtliche Verantwortung für die Durchführung des Forschungsvorhabens beim Leiter und seinen Mitarbeitern liegt.
- Datenschutzrechtliche Aspekte von Forschungsvorhaben werden durch die EK grundsätzlich nur kursorisch geprüft. Dieses Votum ersetzt mithin nicht die Konsultation des zuständigen Datenschutzbeauftragten.
- Laut BDSG § 23 (3) sind personenbezogene Daten zu anonymisieren, sobald der Forschungszweck dies ermöglicht.
- Jedes Forschungsvorhaben, an dem Versuchspersonen beteiligt sind, ist vor Rekrutierung der ersten Versuchsperson in einer öffentlich zugänglichen Datenbank (z.B. Deutsches Register Klinischer Studien) zu registrieren (DvH 2013, § 35).
- Die Zusammensetzung und Arbeitsweise der EK entspricht der ICH-GCP-Leitlinie in der gültigen Fassung.

Anlagen:

Eingereichte Dateien:

Eingereichte Dateien:

- 1) Lebenslauf_Prof_Dr_Dominik_Heider.pdf vom 05.10.2023
- 2) Lebenslauf_Jan_Ruhland.pdf vom 05.10.2023
- 3) Checkliste_Prospektive_Datenerhebung.pdf vom 05.10.2023
- 4) Angaben_Prospektive_Datenerhebung.pdf vom 05.10.2023
- 5) Studienprotokoll.pdf vom 05.10.2023
- 6) Einwilligungserklärung.pdf vom 05.10.2023
- 7) fragebogen.pdf vom 05.10.2023
- 8) Anschreiben_Ethikkommission.pdf vom 05.10.2023
- 9) Gebührenreduzierung.pdf vom 05.10.2023
- 10) Vollmacht.pdf vom 05.10.2023
- 11) Kurzbeschreibung.pdf vom 05.10.2023
- 12) Anschreiben_Ethikkommission.pdf vom 23.10.2023
- 13) Einwilligungserklärung_2.pdf vom 23.10.2023
- 14) Angaben_zur_prospektive_Datenerhebung_2.pdf vom 23.10.2023
- 15) Studienprotokoll_2.pdf vom 23.10.2023
- 16) Vollmacht_2.pdf vom 23.10.2023
- 17) fragebogen_2.pdf vom 23.10.2023
- 18) fragebogen_2.pdf vom 31.10.2023
- 19) Vollmacht_2.pdf vom 31.10.2023
- 20) Studienprotokoll_2.pdf vom 31.10.2023
- 21) Angaben_zur_prospektive_Datenerhebung_2.pdf vom 31.10.2023
- 22) Anschreiben_Ethikkommission_2.pdf vom 31.10.2023
- 23) Einwilligungserklärung_3.pdf vom 31.10.2023