© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712301

Transformer-Based Deep Learning for Multimodal IoT Data Fusion in Smart Healthcare: A Comprehensive Review with Emphasis on Cardiovascular Disease Monitoring and Real-Time Deployment

MUHAMMED KULIYA[1], ZAHARADDEEN SALELE IRO[2]
[1,2]*Department of Computer Science, Federal University Dutse*

Abstract- ***The integration of Internet of Things (IoT) devices and artificial intelligence has revolutionized smart healthcare, enabling continuous monitoring, personalized treatment, and proactive intervention. Cardiovascular disease (CVD) is still one of the main causes of death worldwide, yet we don't have a strong, unified way to monitor it in real time using deep learning. Traditional machine learning methods often fall short; they strive to combine different kinds of medical data, deal with delays in processing, and work reliably in low-resource settings where technology and infrastructure are limited. This review focuses on closing that gap by examining how transformer-based deep learning models can be applied to multimodal IoT data in smart healthcare, with a special emphasis on predicting CVD. We classify and assess the latest transformer architectures based on how they fuse data, their areas of application, and their readiness for real-time use. Our analysis shows that transformer models, with their attention mechanisms and ability to handle information across multiple formats, perform much better than traditional approaches when it comes to combining data from sources like physiological signals, medical imaging, and clinical records. However, we also identify several challenges: high computational demands for edge devices, limited interpretability, a limited multimodal dataset, and infrastructure barriers in under-resourced regions. To address these challenges, we highlight future directions such as creating lightweight transformer models, using privacy-preserving federated learning, and developing unified multimodal pretraining strategies. This review aims to provide a roadmap for building fair, scalable, and low-latency AI solutions for real-time cardiovascular prediction, offering valuable insights for both researchers and healthcare system developers.***

Index Terms- ***Transformer Models, Internet of Things, Multimodal Data Fusion, Cardiovascular Disease Monitoring, Real-Time Deployment***

I. INTRODUCTION

CVD remains the leading cause of death globally, emphasizing the need for continuous and intelligent monitoring systems. As a result of CVD, there were nearly 17.9 million deaths in 2019, a figure representing 32% of worldwide deaths (WHO, 2021). This epidemic requires innovative methodologies for early detection, continuous monitoring, and appropriate intervention. The emergence of the Internet of Things (IoT) leading to Internet of Medical Things (IoMT) has led to a pattern shift in smart healthcare, enabling a unified collection of multimodal physiological data, such as electrocardiograms (ECG), blood pressure (BP), oxygen saturation (SpO_2), and accelerometer readings, through wearable and implantable sensors (Osama et al., 2023). These heterogeneous data streams offer unique opportunities for detailed patient representation; nevertheless, new challenges include integrating such data streams, noise resilience, and real-time processing (Chen et al., 2024).

In unimodal cardiovascular disease analysis, conventional deep learning architectures, particularly CNNs, have proven highly effective (Acharya et al., 2017). However, they struggle to model cross-modal dependencies and long-range sequential patterns rooted in multimodal IoT data.

While transformers were first developed for natural language processing, their adaptability has made them a cornerstone of modern AI. The secret to their success is self-attention, a technique that excels at identifying contextual connections across completely different kinds of data. This capability is a natural fit

for the complex challenge of multimodal fusion (Dosovitskiy et al., 2020). Transformer models for real-time CVD monitoring are a relatively recent phenomenon. They have been somewhat constrained in their computational resources, particularly concerning different data sources and required architectures for edge computing (Noor et al., 2025).

Although many reviews address deep learning for healthcare (Morid et al., 2023) and IoT systems (Li et al., 2024), only very few studies (Saleh et al., 2025) deal specifically with transformer-based multimodal fusion for real-time monitoring of CVD. Same early, late, or hybrid fusion approaches remain fragmented and do not provide standardized benchmarks for CVD applications (Krones et al., 2025). While foundational models like BERT and ViT have demonstrated state-of-the-art performance, their considerable computational complexity and high latency render them unsuitable for deployment on resource-constrained edge devices requiring real-time inference (Madan et al., 2024). Additionally, most studies fail to account for socioeconomic and infrastructural tasks in regions with the highest CVD burden, such as Sub-Saharan Africa (Roth et al., 2020). The limited diversity of publicly available multimodal CVD datasets further obscures generalizability (Wang et al., 2024). While multimodal cardiology frameworks like CardioNet+ achieve 99.1% accuracy and a 99.0% AUC-ROC by integrating ECG/PPG signals with chest X-ray data, they significantly outperform single-modal models in heart failure detection (Adeyi, 2025), surpassing single-modal models and establishing a new benchmark in heart failure identification systems.

II. RELATED WORKS

A significant trend in modern healthcare is the leveraging of IoT-generated multimodal data for advanced analytics in chronic disease management, with notable applications in CVD. The widespread deployment of biomedical sensors, wearables, and imaging platforms has resulted in data ecosystems characterized by high volume, variety, and complexity, presenting both unprecedented opportunities and analytical challenges. Extracting actionable insights demands advanced fusion frameworks capable of handling such data diversity. This section reviews existing literature on multimodal data fusion in smart healthcare, focusing on CVD prediction. It explores IoT's role, sensor technologies, fusion strategies, and the impact of ML, DL, and transformer-based models. Furthermore, it addresses infrastructure for real-time monitoring and discusses challenges in building scalable, interpretable fusion systems, highlighting gaps and future research directions in cardiovascular healthcare applications.

A. Multimodal IoT in Smart Healthcare

Unimodal data collects information from a **single source** rather than combining multiple data streams. In implies that, models rely on only one category of modality to make predictions. Zhong et al. (2022) in their study highlights the limitations of unimodal models in predicting antenatal depression. While unimodal methods depict certain patterns, their accuracy, sensitivity, and reliability are limited. In contrast to multimodal methods, they fail to reflect the complex cross-modal interactions affecting maternal mental health, making them less effective and likely to miss important predictive signals essential for early detection and intervention (Zhong et al., 2022).

Multimodal IoT refers to the combined use of different types of data, like physiological signals, environmental sensors, medical images, and EHRs to improve how we monitor patients and make diagnoses. In recent years, the integration of these multimodal IoT technologies into smart healthcare systems has drawn a lot of attention, as it helps deliver real-time, personalized, and proactive medical care.Initial efforts were focused on unimodality IoT-based monitoring systems for health assessment, such as wearable ECG monitors or glucose sensors (Mahmmod et al., 2024). The unimodality monitoring approaches presented uneven views of the health state of a patient, and to address this problem, researchers began to integrate multiple sensor modalities with different matching health data.

Multimodal fusion of heterogeneous streams of physiological data from wearables (ECG, photoplethysmography (PPG)), ambient sensor signals, and implantable devices is known to introduce complementary aspects of cardiovascular

activity and thus enable better patient understanding that it would not otherwise be possible to achieve using individual modalities (Ahmad et al., 2025; Moon et al., 2023). ECG can monitor direct cardiac activity as a subset of the heart's electrical signal. Still, PPG shows peripheral hemodynamics, and accelerometers capture motion variability, thus further improving robustness against noise and uncertainty (Xing et al., 2025).

Recent studies have adopted ML and DL models for effective multimodal fusion (Chen et al., 2025). Cretu et al. (2023) demonstrated that combining ECG, arterial blood pressure (ABP), and central venous pressure (CVP) signals significantly improved the accuracy of arrhythmia detection, with a ResNet50 model achieving 99.58% accuracy across five arrhythmia classes. Likewise, Wang et al. (2023) recent work revealed that healthcare systems are integrating AI, big data, and wearable IoT technologies, and highlighted improved health management and disease prevention through the combination of physiological signal monitoring, personalized elderly care, and EHRs.

In many studies, fusion approaches differ between studies. Some may use early fusion (with pre-processed raw data streams combined before feature extraction), while others may apply late fusion, where predictions from one or more individual modalities are merged. Hybrid approaches, which combine both feature and decision-level integration, are also emerging to optimize performance (Kulasekara et al., 2025). These data fusion techniques frequently leverage deep learning architectures, including CNNs, RNNs, and, more recently, transformer-based models, which are particularly adept at capturing long-range dependencies across different modalities,

B. IoT in Cardiovascular Healthcare

CVD is responsible for 17.9 million yearly deaths worldwide, representing 32% of global mortality (WHO, 2021). Developing nations are confronting a growing burden of CVD, driven by rapid epidemiological transitions and constrained healthcare access. The IoT offers a promising approach to mitigate these challenges by facilitating remote patient monitoring. Wearable sensors, which track parameters such as ECG, blood pressure, and pulse, enable the continuous collection of real-time physiological data. The multimodal data collected is analyzed to improve early detection and risk stratification beyond traditional unimodal systems. IoT solutions, such as Predictis, demonstrate potential for scalable preventive care, particularly in resource-limited settings, by combining affordable sensors with user-friendly mobile interfaces for proactive CVD management (Islam et al., 2023).

Unimodal IoT systems focus on acquiring single-parameter physiological data for cardiovascular monitoring, typically concentrating on either ECG or echocardiogram (Echo) signals. Single-signal monitoring systems have important limitations in CVD care. They track just one type of body signal, fails to account for the broader context of heart health (Yan et al., 2022). Consequently, in the absence of complementary data such as blood pressure or oxygen saturation, these methods cannot provide a comprehensive assessment of the risks associated with complex cardiac conditions. While effective for detecting specific anomalies like arrhythmias, they are often limited in their capacity to identify broader cardiovascular risk factors. Likewise, these systems are more likely to give wrong readings due to body movements or technical errors (Wang et al., 2024).

To overcome the single-signal constraint, multimodal IoT systems are used that combine multiple health data, such as ECG, PPG, blood pressure, oxygen saturation, breathing rate, etc., which offer healthcare professionals a more comprehensive view of heart health (John et al., 2024; Boikanyo, et al., 2023**).** The fusion of multimodal data from complementary sensors enhances the accuracy and robustness of smart healthcare systems, thereby improving their overall reliability. Multimode data fusion greatly reduces noise and accurately finds optimal impact for achieving high efficiency in the health sector (Kern, 2025). In addition, it provides for real-time processing and contextual understanding for systems of increasing complexity (e. g. medical diagnosis and autonomous systems).

C. Biomedical Sensors for Multimodal Data Fusion in CVD Monitoring

The integration of multimodal data has significantly enhanced the monitoring of CVD, leading to more accurate, reliable, and clinically actionable patient assessments. This progress is largely driven by advances in biosensor technology, which can capture complementary physiological signals. Such sensors facilitate continuous monitoring and improved risk stratification in both clinical and ambulatory settings. This section examines the principal sensor modalities employed in these multimodal systems, detailing their physiological targets, key technological features, and representative applications in contemporary research.

- ECG Sensors

ECG is the cornerstone of cardiac monitoring, providing critical insights into the heart's electrical activity. Consequently, ECG sensors are essential for identifying arrhythmias, ischemia, myocardial infarction, and other cardiac pathologies. Chester straps, patch-type monitors, and wearable ECG equipment can be used to collect data continuously and provide a wide range of patient mobility and flexibility for long-term cardiac surveillance. New devices that combine ECG signals with additional signal monitoring techniques such as PPG and accelerometry have been developed and evaluated for improved arrhythmia detection and stress monitoring accuracy (Alimbayeva et al., 2024).

- PPG Sensors

Kim and Baek (2023) review the current state of PPG technology for wearable devices, including its non-invasive use in a range of applications, including monitoring heart rate, oxygen saturation, blood pressure, sleep quality, and stress. They review technology developments in small multi-wavelength sensors and low-power consumption systems and identify critical challenges that include motion artifacts, measurement accuracy, skin tone variability, and battery life limitations.

Kim and Baek (2023) stated that further research in PPG will focus on further improving accuracy in 24-hour continuous monitoring, developing novel health parameters, improving cuffless blood pressure monitoring and glucose monitoring, and analysis of stress and sleep. Further progress in multi-wavelength sensors, adaptive algorithms, and real-world validation studies in clinical applications is critical for further strengthening the reliability of PPG and its applications in wearable healthcare technology.

- Impedance Cardiography (ICG) Sensors

According to Mansouri et al. (2022), ICG offers a non-invasive and cost-effective method for diagnosing CVDs. This technique measures thoracic bioimpedance to estimate real-time hemodynamic parameters, including cardiac output, stroke volume, and arterial compliance. Consequently, ICG has diagnostic utility for a range of conditions, such as valvular heart disease, hypertension, arrhythmias, vascular disorders, heart failure, and Cushing's syndrome. The authors state that the use of a large-scale ICG signal database would improve the automatic diagnosis of CVDs by artificial intelligence and in other ways address the existing issues, such as a lack of signals, voltage variability between electrodes, and smaller-scale clinical trials. A fusion system of ICG data in AI for automatic cardiovascular disease diagnosis needs to be explored (Mansouri et al., 2022). The authors also emphasize the need for a comprehensive database of ICG signals to enhance the diagnostic accuracy of machine learning models, mitigate the challenges of signal variability, and advance the development of automated CVD detection systems.

- Blood Pressure (BP) Sensors

Islam et al. (2023) describe BP sensors as a core component of a wearable IoT-based health monitoring system named Predictis. It involves an automatic blood pressure monitor on a wrist wearable type JZK-003 for real-time BP measurement, and its data is transmitted via Bluetooth to a mobile app for continuous cardiovascular monitoring and CVD risk level prediction. Data from the BP sensor in real-time can also assist in an accurate and timely assessment of heart health conditions (Islam et al. 2023). The

authors emphasize that future BP sensor systems should not only provide accurate measurements of BP in real-time, but also implement algorithms from cloud-based platforms to support continuous monitoring, risk prediction, emergency warning, and mobile application interfacing.

- Respiratory Rate and Thoracic Motion Sensors

Respiratory parameters are closely linked with cardiac function. Sensors for instance respiratory inductance plethysmography bands, strain gauges, and accelerometers, when placed on the chest, are used to take respiratory rate and tidal volume. These sensors are valuable in conditions such as heart failure and sleep apnea, where ventilation-perfusion mismatch and sympathetic overactivity are prominent (Ceccarelli et al., 2022).

The combination of respiratory signals with cardiovascular data has led to improved models for detecting sleep-disordered breathing and assessing cardiorespiratory coupling. For example, smart garments that integrate ECG, PPG, and respiratory motion sensors have been used to monitor nocturnal events and assess autonomic dysfunction in patients with heart disease (Lu et al., 2024).

D. Data Transmission and Integration Mechanisms for Multimodal Data Fusion

Multimodal data fusion systems integrate information from diverse sources to provide a more comprehensive understanding, which is crucial in various applications like healthcare (Abdar et al., 2023). The process involves several mechanisms for data transmission and integration to overcome the limitations of single-modal data (Abdar et al., 2023).

- Data Transmission Mechanisms

Multimodal data, which encompasses both structured and unstructured formats, is generated in massive volumes daily by a diverse array of sensors and systems (Ahmad et al., 2025). A prominent example is the IoMT, where networks of sensors continuously collect various health metrics, including vital signs, physical activity levels, and ECG readings (Adedinsewo, 2023). This data is then transmitted through various layers and protocols:

- Sensor Layer: This foundational layer in IoMT systems collects data from patients using sensors, controllers, and actuators (Ahmad et al., 2025). It includes a data-entry sublayer for signal acquisition, utilizing techniques like General Packet Radio Service (GPRS), Radio Frequency Identification (RFID), and graphic codes (Ahmad et al., 2025).
- Transmission Technologies: Collected data is securely transmitted to a central location (e.g., cloud server, hospital data center) (Adedinsewo et al., 2023). Short-range data transmission methods include Bluetooth, Bluetooth Low Energy (BLE), Wi-Fi, and Zigbee (Koulouras et al., 2025). Long-range communication approaches, such as LoRa, Sigfox, 4G, and 5G, are also employed (Al-Shareeda et al., 2023). Ethernet offers robust and high-speed wired transmission for applications that require high bandwidth.
- Gateway Layer (Fog/Edge Layer): This layer enables real-time data transfer and data preparation, combining different networks, data warehouses, and data description formats. Edge computing frequently performs data preparation at this level, closer to the data sources (Yıldırım et al., 2025).
- Cloud Layer: Large medical and healthcare systems integrate with the cloud for daily operations, including storing patient data and processing updated medical samples (Banimfreg, 2023).
- Layered Architecture: IoT communication, including multimodal data transmission, is organized into a layered architecture like the OSI model, which ensures efficient data exchange and proper handling across different protocols. Each layer has a specific task, from physical connection (Physical Layer) to error-free transmission (Data Link Layer), routing (Network Layer), reliable delivery (Transport Layer), session management (Session Layer), data format translation (Presentation Layer), and user application interface (Application Layer) (Gupta et al., 2024).
- Specialized Protocols: Certain protocols are used for certain data types and applications, such as OBD2/CAN-BUS for vehicle diagnostic data and OPC UA for secure industrial data exchange (Henke, 2022).

- Data Integration Mechanisms

Data integration, frequently termed data fusion, addresses the challenge of insufficient or noisy single-source data by combining information from multiple modalities. The objective is to leverage the complementary and redundant information inherent across these diverse sources. The methodologies for this integration have been profoundly transformed by deep learning, marking a paradigm shift from reliance on hand-crafted feature engineering to automated, learned representation. Fusion strategies are generally categorized by the stage at which data from different modalities are combined within the processing pipeline:

- Early Fusion (Data-Level Fusion)

In early fusion, raw or minimally pre-processed data from multiple modalities are combined at the input level and fed into a single processing model. This approach maintains the original information from each modality and decreases computational costs by means of a single encoder, letting the model to learn cross-modal relationships from low-level features. However, it can result in very high input dimensions with multiple modalities and is best suited for homogeneous data or a limited number of modalities, as it may fail to capture relationships that emerge at higher abstraction levels (Kulasekara et al., 2025).

- Intermediate fusion (Feature-level fusion)

Intermediate fusion, or feature-level fusion, involves processing data from each modality separately to extract feature vectors, which are then combined within the network before making a final decision. This method allows flexibility in how and when features are fused, enabling precise modeling of relationships and ensuring heterogeneous data are transformed into comparable feature vectors, making it robust to missing modalities and dimensional imbalances. It includes marginal fusion, where features are concatenated before classification, and joint fusion, where additional layers learn abstract cross-modality interactions (Guarrasi et al., 2025).

- Late Fusion (Decision-Level Fusion)

Late Fusion, or decision-level fusion, is a multimodal integration strategy where separate models are trained independently on distinct data modalities. The final decision is derived by aggregating the outputs of these models through techniques such as weighted averaging or majority voting. This approach allows each model to be fine-tuned for its specific data type, which can lead to complementary, uncorrelated errors and is simple to implement, even combining deep and shallow learning methods. However, it cannot capture interactions between features at the data or feature level, as fusion only happens at the decision stage (Kulasekara et al., 2025).

- Hybrid Fusion

Hybrid fusion systems merge components of early, intermediate, and late fusion, dynamically selecting the suitable fusion level based on task requirements and environmental settings (Shaik et al., 2024). Beyond traditional methods, deep learning has introduced fine-grained techniques like encoder-decoder fusion, which maps multimodal data into latent spaces for flexible prediction. Attention-based fusion selectively weighs inputs through self-attention and cross-attention, effectively modeling dependencies within and across modalities, as seen in Transformer architectures. Graph Neural Networks (GNNs) provide a natural framework for modeling relational multimodal data by representing it within a unified graph structure. To complement, Generative Neural Networks (GenNNs) can be employed to synthesize missing data modalities or to enforce semantic consistency across them. The choice of an optimal fusion strategy is contingent upon the data characteristics, specific application requirements, and the critical trade-offs between model accuracy, robustness, and computational efficiency (Shaik et al., 2024).

- Significance of Multimodal Data Fusion in CVD Prediction

The complexity of CVD arises from its multifactorial nature, involving genetic, physiological, behavioral, and environmental components (Valeria et al., 2024). Thus, traditional predictive models relying on unimodal data regularly fall short in capturing the complex interrelationships that bring about CVD onset and progression. In multimodal data fusion, the integration of heterogeneous data sources such as clinical records, imaging, genomics, and wearable sensor data has emerged as a pivotal strategy for

enhancing the precision and robustness of predictive models (Li et al., 2024).

Multimodal data fusion enables a comprehensive representation of patient health by integrating complementary data types. EHRs provide detailed information on patient history, diagnoses, and medications, while medical imaging techniques such as echocardiography or CT angiography deliver spatial and morphological characterization of cardiac structure and function **(**Zhou et al., 2023**)**. Integrating these with genomic data can illuminate genetic predispositions to atherosclerosis or cardiomyopathy, while wearable sensor data can capture real-time physiological signals like heart rate variability, physical activity, and sleep patterns. The integration of diverse data streams through multimodal fusion yields patient profiles of greater granularity and contextual richness, which in turn enhances predictive performance. Concurrently, advances in ML and DL have accelerated the adoption of these multimodal approaches. Techniques such as deep learning are particularly well-suited for this task, as they can model complex, non-linear relationships across different data modalities. For instance, Solares et al. (2020) demonstrated that DL models, capable of processing large-scale, multimodal, and sequential EHR data, significantly outperform traditional statistical models in clinical risk prediction. This superior performance is largely attributable to the ability of deep learning to automatically learn intricate patterns and interactions directly from raw data. Similarly, the work of Lu et al. (2024) shows that deep learning models employing multimodal data fusion significantly outperform traditional methods. Their approach effectively learns both shared and subtype-specific patient representations. By integrating knowledge graphs, the model not only enhances its interpretability but also mitigates data scarcity through a shared-private feature learning framework. This approach improves clinical prediction tasks, such as disease outcome predictions, even in few-shot and zero-shot scenarios (Krones et al., 2025).

The integration of multimodal data significantly improves model interpretability and clinical utility. By combining structured information, such as lab results, with unstructured data, like clinical notes, models can achieve a more holistic representation of disease pathology. This comprehensive approach facilitates the discovery of novel biomarkers and the identification of complex risk factors (Shaik et al., 2024). Additionally, the fusion of temporally combined data, such as longitudinal EHRs and constant monitoring from wearable devices, aids the development of dynamic prediction models that can adjust to changes in patient health condition over time.

- Transformer Models for Multimodal Data Fusion in Healthcare

Transformer-based models have rapidly proven their ability to design sophisticated multimodal data fusion algorithms capable of representing complex relationships across heterogeneous data streams. In the context of healthcare, in which data are usually acquired from diverse modalities such as time series physiological signals, imaging, and clinical notes, transformers provide a flexible and scalable method required for disease diagnosis and monitoring.

Transformers bring multimodal cardiovascular data fusion to a new level of self-attention by relying on attention mechanisms across heterogeneous streams of various data, including ECG, PPG, imaging, and clinical notes (Noor et al., 2025). The core innovation in transformer-based fusion comprises:

- Cross-modal attention: Cross-modal attention **is** a neural mechanism in transformer architectures that supports dynamic interaction between diverse data sources **(e.g.,** ECG signals and clinical text) by calculating relevance scores between modalities. Cross-modal attention synchronizes asynchronous data streams by dynamically aligning temporal events (Zhu et al., 2024). It identifies clinically significant inter-modal relationships by learning latent connections and it suppresses noisy modalities (e.g., ignoring motion-corrupted PPG during exercise) using gated attention.
- Lightweight Architectures for Edge Deployment: Designing lightweight transformer architectures for real-time, on-device healthcare monitoring using the reduced computational burden of traditional models (like BERT). With the help of knowledge distillation and pruning techniques,

models like DistilBERT can achieve a very high-performance level with fewer parameters. Cardiovascular care frameworks like CardioNet+ have often been built using cloud processing, but now with lightweight adaptations that can run on edge devices (Psomakelis et al., 2023). That allows continuous CVD monitoring without internet access, which was extremely important in remote or even rural settings.

- Transformer-based models enable the combination of structured data (ECG, vital signs) and unstructured data (clinical notes, photos, etc.) via attention mechanisms to a larger understanding and hence the predictive ability (Madan et al., 2024).

- Transformer-based Model for Categorical Data

Categorical and integer data are essential elements of structured datasets in healthcare analytics. Categorical variables represent discrete, non-numerical groups, such as diagnostic classifications or patient demographics. In contrast, integer variables encompass numerical counts or measurements, including age or blood pressure readings. Effectively modeling these data types requires specialized techniques to capture their distinct patterns and relationships, ensuring accurate predictions in clinical and epidemiological studies. The key models include**:**

- Tab Transformer encodes categorical columns using embedding layers and applies self-attention mechanisms to model feature interactions within tabular data. This approach provides a more robust representation of categorical variables, which enhances model performance in both classification and regression tasks. It has been effectively applied in domains like finance, healthcare, and retail (Alam et al., 2023).
- FT-Transformer introduces a framework where structured tabular data is tokenized into feature tokens, which are processed using a standard Transformer encoder. This architecture generalizes well across diverse tabular machine learning tasks by learning complex feature relationships. It demonstrates strong performance benchmarks in general-purpose applications across various industries (Gutheil & Donsa et al, 2022).
- SAINT leverages both intra-feature attention (capturing dependencies among features) and inter-sample attention (modeling relationships across different data instances) in tabular datasets. This dual-attention mechanism enhances predictive performance on classification and regression problems. SAINT has shown notable improvements in tasks involving categorical data representations (Gutheil & Donsa et al., 2022).
- TabNet applies a sequential attention mechanism on feature subsets, rather than using a full Transformer architecture, to maintain interpretability while processing tabular data. The model's attention mechanism facilitates feature selection by identifying the most salient variables for tasks such as risk prediction, fraud detection, and clinical decision support. Consequently, TabNet has become a model of choice in domains that demand not only high accuracy but also transparent, explainable decision-making (Alam et al., 2023).
- Med-BERT adapts the BERT architecture customized to process structured diagnosis codes from Electronic Health Records (EHRs), such as sequences of International Classification of Diseases (ICD) codes. By leveraging pretraining on large-scale medical datasets, it achieves state-of-the-art performance in predictive tasks, including chronic disease prediction and patient outcome modeling. This model enhances EHR-based clinical analytics through context-aware encoding of medical codes (Rasmy et al., 2021).
- BEHRT employs a time-aware Transformer model to capture patient health trajectories through sequences of diagnosis codes. By integrating temporal information, it models disease progression and patient history with greater clinical relevance. BEHRT has proven effective in longitudinal healthcare applications for early disease detection (Li et al., 2024).

- Retain-Transformer is designed for longitudinal health records, incorporating interpretability-focused attention mechanisms that highlight influential clinical events over time. This model facilitates risk prediction tasks, such as forecasting heart failure and other adverse outcomes, by making temporal relationships in EHRs transparent. It bridges performance with interpretability in clinical decision-making (Lentzen et al., 2023).

- Transformer-Based Models in Medical Imaging
Transformer-based models have demonstrated significant potential in analyzing complex visual patterns within medical imaging, including MRI, CT, and X-rays. These images contain rich spatial and contextual information that is critical for accurate diagnosis and effective treatment planning. The key models include**:**

  - TransUNet fuses a CNN-based encoder with a transformer-based decoder to model both local texture details and global anatomical structures in medical images. The hybrid architecture yields superior segmentation accuracy, especially in delineating complex boundaries. It has been effectively applied to CT and MRI image segmentation tasks in clinical settings (Chen et al., 2021).
  - UNETR employs a pure transformer encoder connected to a CNN decoder via skip connections to capture long-range dependencies in 3D volumetric data. This design achieves precise segmentation of complex anatomical structures in MRI scans by building upon UNETR, a model with established utility in 3D medical imaging applications (Hatamizadeh et al., 2022).
  - Swin UNet incorporates Swin Transformer blocks for hierarchical feature extraction, allowing effective multiscale representation of medical images. This method enhances segmentation performance by modeling both local and global spatial relations through shifted window attention. It has shown significant success in organ and tumor segmentation tasks using CT and MRI data (Cao et al., 2022).
  - MedT **i**ntroduces Gated Transformer Units (GTUs) to enhance the segmentation of medical images, especially when annotated data is scarce. By leveraging an attention mechanism to focus on clinically relevant regions, MedT achieves high diagnostic accuracy even in data-scarce clinical environments. It has been applied to segmentation tasks on CT and X-ray datasets (Valanarasu et al., 2021).
  - DINO-ViT leverages self-distillation-based Vision Transformers to learn robust feature representations without labeled data, enhancing classification performance in medical imaging. The proposed self-supervised framework offers a significant reduction in the need for expensively annotated data, while preserving a level of diagnostic accuracy comparable to supervised methods. DINO-ViT has been utilized in classification tasks involving X-ray and CT images (Anand et al., 2023).
  - BioViL-T aligns chest X-ray images with respective radiology text reports using a vision-language transformer trained with contrastive learning. This approach improves multimodal understanding, facilitating automated report generation and image-text retrieval. BioViL-T has been successfully applied to radiology report comprehension and multimodal diagnostic tasks (Bannur et al., 2023).
  - SwinIR-Med adapts Swin Transformer architectures for image enhancement tasks, focusing on super-resolution and denoising in low-quality medical images. The model effectively restores details in noisy MRI and PET scans, improving image clarity for clinical interpretation. SwinIR-Med has been used in scenarios requiring high-fidelity reconstruction of degraded medical images (Puttagunta et al., 2022).
  - ViT-Medical fine-tunes Vision Transformers specifically for medical image classification and lesion detection, optimizing them for clinical datasets. This adaptation enables accurate identification of disease markers in X-ray and CT scans. ViT-Medical has been

applied across various diagnostic imaging tasks in healthcare (Manzari et al., 2023).

- Regressive Vision Transformer (RVT) combines global self-attention from Vision Transformers with localized attention approaches for enhancing feature extraction in radiology images. The hybrid approach enhances disease classification accuracy by leveraging both coarse and fine-grained patterns. RVT has been applied to chest X-ray and CT-based disease classification tasks (Li and Zhang, 2024).

E. Machine Learning and Deep Learning Models for CVD Prediction

The persistent global burden of CVD mortality has motivated the exploration of advanced computational methods. This has led to a progression from traditional ML to DL, and more recently, to hybrid models that integrate multiple approaches to harness their complementary advantages.

- Traditional ML

  - Logistic Regression (LR)

    Despite their assumption of linearity, which is an inherent limitation, logistic regression (LR) models remain valuable in CVD risk prediction due to their computational efficiency and interpretability. However, since linear relationships between risk factors are assumed to be constant throughout the model, LR models struggle to generalize among a sample of individual risk factors such as CVD to a much wider population and thus tend to underestimate risk in people younger, females, and minorities. Due to these limitations, more sophisticated and flexible machine learning models have been developed that can capture the dynamic and multifactorial risk profile more precisely (Kasartzian and Tsiampalis, 2025**).**

  - Random Forest (RF)

    Yang et al. (2024) in their paper adds that Random Forest is a promising classifier for CVD achieving 91% accuracy on the Long Beach VA dataset and 90% AUC after applying the proposed data balancing technique. RF is found to be superior to other classifiers due to its ability to reduce overfitting and the difficulties in handling imbalanced datasets making it suitable for clinical applications. The model's interpretability, achieved through SHAP values, aligns with established clinical expertise, thereby fostering greater trust in its utility for clinical decision-making. (Yang et al., 2024).

  - Deep learning (DL)

    Deep learning has revolutionized the medical imaging industry by giving rise to automated and high-quality feature extraction and interpretation with significantly better accuracy compared to existing methods for diseases detection, segmentation, and classification in more challenging image data sets such as MRI, CT scan, and ultrasound.

  - Convolutional Neural Network (CNN)

    There is increasing importance of CNNs for the analysis of medical images for CVD, with particular emphasis on applications such as classification, segmentation and detection (Jia et al., 2024). Common CNN models like ResNet and U-Net are commonly used for the analysis of CT and MRI images, which are at the center of CVD research (Jia et al., 2024). The authors acknowledge challenges, including the cost of data annotation and data privacy concerns. Nonetheless, they emphasize the potential of Convolutional Neural Networks (CNNs) to enhance diagnostic accuracy and improve workflow efficiency. Emerging areas include multimodal learning and federated learning to overcome data limitations (Jia et al., 2024).

  - Long Short-Term Memory (LSTM)

    LSTM networks, a specialized variant of Recurrent Neural Networks (RNNs), were developed to model long-term temporal dependencies in sequential data, such as electrocardiogram (ECG) signals. They overcome the vanishing gradient problem via memory cells and gating mechanisms, enabling retention of temporal patterns across

extended heartbeat sequences. This makes LSTMs effective for detecting arrhythmias dependent on multi-beat irregularities (e.g., Atrial Fibrillation). However, these models are computationally demanding (e.g., standard RNNs or MLPs) (Ansari et al., 2023).

- Hybrid Model

Hybrid models represent a growing trend in which multimodal data are processed by integrating complementary architectures to achieve superior predictive performance.

- CNN-LSTM model

CNN-LSTM model incorporates convolutional feature extraction layers with LSTM layers to capture temporal dependencies. Sudha and Kumar (2023) in their heart disease prediction designed an architecture consisting of 5 convolutional and pooling layers followed by LSTM and fully connected layers with Softmax activation. On the basis of UCI heart disease tabular data, the missing values are normalized with z-score and features are selected by the SVM weighting method. The authors train the model with the Adam optimizer over 200 epochs with a learning rate of 0. 001% and obtain 89% accuracy, 81% sensitivity and 93% specificity. It outperforms traditional classifiers due to their dynamic dimension of the data.

- CNN-Transformer

Hybrid Vision Transformer (HVT) architectures integrate the complementary strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). They typically leverage CNNs to extract fine-grained local features and employ the self-attention mechanisms of ViTs to capture long-range global dependencies. To capture complementary information at both local and global scales, studies have employed various integration strategies, notably sequential, parallel, and hierarchical integration. In general, HVTs outperform standalone CNNs and ViTs in image recognition and segmentation tasks in limited datasets. In addition, the technique convolutional token embedding enhances efficient computation, for example by decreasing computational cost without compromising accuracy (Sagheer et al., 2025).

- Graph Neural Networks (GNNs)

The application of Graph Neural Networks (GNNs) offers a robust paradigm for analyzing healthcare data, which is inherently relational. By representing medical entities as nodes and their interactions as edges, GNNs can directly model the complex dependencies within such data. Architecturally, GNNs are categorized into recurrent, spatial, and spectral approaches, which support fundamental tasks including node-level, link-level, and graph-level prediction. These models consistently surpass traditional machine learning methods, primarily due to their superior handling of data heterogeneity, temporal dynamics, and sparsity. Consequently, the proficiency of GNNs in modeling dynamic processes, fusing multi-modal inputs, and offering explainable predictions establishes them as a pivotal technology for advancing healthcare, particularly in the realms of disease prognostication, drug repurposing, personalized treatment planning, and enhanced clinical decision-support systems (Paul et al., 2024).

- Autoencoder–Random Forest

Hybrid autoencoder–random forest architecture primarily within the context of anomaly detection. This hybrid approach employs autoencoders for efficient feature extraction and dimensionality reduction, enabling the learning of compact and semantically meaningful representations of the input data. The extracted features are subsequently fed into a Random Forest classifier, which demonstrates strong performance in handling classification tasks, particularly in scenarios involving imbalanced or complex datasets (Berahmand et al., 2024). This hybrid model leverages the unsupervised feature learning of autoencoders and the strong classification capability of random

forests to achieve higher detection accuracy and better generalization across diverse datasets (Berahmand et al., 2024).

- CNN-SVM
  Hybrid CNN-SVM architecture integrates CNNs for hierarchical feature extraction with support vector machines (SVMs) for robust classification of cancer types using high-dimensional RNA-Seq data. Among the evaluated models, the Hybrid-CNN-SVM architecture, incorporating parallel convolutional layers, achieved the highest accuracy (96%), surpassing the performance of the standalone CNN and SVM models. This hybrid approach enhances generalization through the SVM's structural risk minimization principle and improves robustness to noise, effectively addressing critical challenges in biomedical data analysis. The results demonstrate the hybrid model's effectiveness for complex classification tasks in genomics and precision medicine applications (Nejad, 2025).

F. Application Domains of Multimodal Data Fusion and IoT in Healthcare

It is vital to emphasize the contributions that IoT, combined with multimodal data fusion, as a whole, has made towards modern healthcare by providing an ability for more rapid acquisition, transmission, and analysis of biomedical data. The integration of these factors has resulted in enhanced predictive accuracy, personalized healthcare delivery, and improved anticipatory care. The scope of multimodal data fusion and IoT technologies is extensive, encompassing both clinical and non-clinical domains, including remote patient monitoring, chronic disease management, medical imaging, critical care, psychological assessment, and rehabilitation services. In this section, this study explores the key medical domains in which multimodal data fusion and the IoT have exerted significant influence.

- Remote Patient Monitoring and Telemedicine

Remote patient monitoring (RPM) represents a key application domain, particularly for aging populations and individuals with chronic conditions such as cardiovascular diseases, diabetes, and respiratory disorders. IoT-based wearable sensors and mobile health platforms allow continuous tracking of physiological signals, heart rate, blood pressure, oxygen saturation, body temperature, and ECG data, which are fused with contextual and environmental information to create holistic health profiles (Boikanyo et al., 2023).

The integration of NLP of physician–patient conversations with facial expression and physiological signal analysis enhances remote consultations by enabling the capture of both verbal and non-verbal cues. This multimodal approach enables deeper patient understanding, supports mental health assessments, and improves diagnostic accuracy during telemedicine interactions by analyzing emotions, stress, and physiological states in real-time (Farrokhi et al., 2024).

Chronic Disease Management

Ahmad et al. (2025) provides a review of emerging rapid detection methods for monitoring CVDs, emphasizing wearable sensors (ECG, PPG) and point-of-care (POC) technologies for real-time biomarker detection. While AI-driven analytics have demonstrated the potential to enhance diagnostic accuracy, persistent challenges concerning signal integrity and data security continue to pose limitations. Future directions should prioritize the integration of personalized medicine approaches, interdisciplinary collaboration, and technological innovations to optimize chronic disease management.

Medical Imaging and Radiomics

Assen et al. (2023) highlight the use of imaging and AI-based fusion modeling techniques in cardiac care. Imaging modalities such as CT, CMR, echocardiography, and nuclear imaging provide important biomarkers such as coronary calcium, plaque volume, and patient age to help cardiovascular risk prediction and patients' personalized treatment. AI provides automatic extraction and analysis of imaging features to improve both clinical and general assessment accuracy and efficiency. Advances in medical imaging technologies have significantly enhanced capabilities in early heart failure detection, digital heart twin construction for personalized ablation planning, and predictive modeling of drug responses.

The authors further state that fusing clinical and imaging data can pointedly improve risk prediction and personalization in cardiovascular disease, and combining it with AI models will result in a higher diagnostic performance, easier treatment planning, and greater prognostic insight, particularly for cardiovascular conditions, including coronary artery disease and heart failure.

Intensive Care and Emergency Medicine

The intensive care units (ICU) high-stakes nature demands a shift from reactive to proactive care, achievable only through multimodal IoT systems. The fusion of wearable, visual, and environmental sensing modalities enables continuous, high-resolution monitoring of functional and behavioral metrics, such as mobility, pain, and sleep, thus addressing key limitations in current healthcare monitoring practices. Achieving success depends on addressing privacy challenges, ensuring interoperability, and demonstrating real-world effectiveness through rigorous validation. Berikol et al. (2025) in their study demonstrate how multimodal AI in emergency medicine combines imaging, EHRs, and physiological data to improve diagnostics. Even though standardization challenges persist, these methods enable holistic evaluations.

Mental Health and Behavioral Analysis

Guo et al. (2022) propose the CASTLE framework, which utilizes multimodal data fusion to assess students' mental health status by integrating social life, academic performance, physical, and demographic features. The framework leverages representation learning in conjunction with a multi-view embedding algorithm for social networks and a deep neural network (DNN) for detection. The experimental results demonstrate its efficacy in identifying mental health issues while effectively addressing challenges such as data heterogeneity and label imbalance. The authors further analyze behavior via multi-view social networks (friendship, advice-sharing) using the MOON algorithm to detect mental health risks. It links social patterns (e.g., isolation, cooperation) to psychological states but notes limitations like static data and self-report bias.

Rehabilitation and Assistive Technologies

Multimodal data fusion in assistive healthcare technologies to enhance functionality for individuals with neurological disabilities. It integrates brain–computer interfaces (BCIs), AI-driven devices, and sensor-based systems to enhance mobility, communication, and cognitive assistance. For instance, the integration of BCIs with virtual reality (VR) facilitates real-time monitoring of cognitive load, while AI algorithms dynamically adapt assistive technologies (ATs) based on users' behavioral and physiological data. Challenges include usability and ethical concerns, but multimodal approaches promise personalized, adaptive solutions (Bonanno et al., 2025).

Raj and Kos (2024) in their article highlight that multimodal sensor fusion is pivotal in assistive robotics, enhancing perception and interaction by integrating data from LiDAR, IMUs, EMG, and vision sensors. This fusion improves environmental awareness, intention recognition, and adaptive control, enabling safer Human-Robot Interaction (HRI). Combining infrared and IMU data aids navigation for the visually impaired, while EMG and vision enable responsive prosthetics. Future advancements in deep learning–based fusion are anticipated to further enhance real-time adaptability and personalized user assistance.

Challenges of Multimodal Data Fusion Systems

Multimodal data fusion offers significant advantages in healthcare, including enhanced diagnostic accuracy and comprehensive access to diverse medical information. Its application and use, however, present several significant challenges. The challenges of multimodal data fusion systems include:

- Data Quality and Interoperability: One of the challenging problems is meeting the quality and interoperability of data that originates from different sources in healthcare, which requires the creation of quality data standards and comprehensive interoperability frameworks (Kumar et al., 2024).
- Privacy and security: Protection of sensitive patient data, obtained from multiple sources, is critical; such protection should encompass

mechanisms like encryption and secure storage techniques, and privacy-preserving techniques, as well as continuous monitoring and auditing for ensuring data integrity and confidentiality of the data (Shaik et al., 2024).

- Data Processing and Analysis: Scalability of systems has become an important and difficult aspect in this area, and therefore, the fusion of data requires the development of different approaches of ML and AI, as well as a scalable data processing and real-time analytics infrastructure (Shaik et al., 2024).
- Clinical Integration and Adoption: Effectively integrating multimodal fusion into current clinical practice demands the effective involvement of healthcare professionals. It also necessitates the design of user-friendly interfaces, the provision of adequate user training, and the integration of these emerging technologies into existing medical systems (AL-Mosawi and Al-Shammari, 2024).
- Ethics considerations: Privacy, autonomy, and fairness concerning patients should be important ethical considerations; this includes obtaining informed consent, establishing ownership of data, detailed governance policies, and rigorously remediating any biases, whether in the data or the algorithm used (Shaik et al., 2023).
- Interpretation of Results: The complexity of multimodal fusion outcomes presents interpretive challenges; therefore, employing visual analytics, explainable AI approaches, and robust clinical validation is vital to ensure meaningful and clinically relevant insights (Shaik et al., 2024).

Though multimodal data fusion is essential for a holistic understanding of patient health in smart healthcare, it requires careful consideration and ongoing research to address its complex technical and ethical aspects.

Real-Time Applications and Challenges of CVD Monitoring

Real-time monitoring systems have become a crucial tool for early diagnosis and intervention of diseases such as arrhythmias, myocardial infarction, and hypertension (WHO, 2021). Using wearable sensor technologies, mobile applications, and cloud computing has significantly improved CVD risk through the collection, analysis, and acting on the physiological signals to decrease mortality and improve medical results (Boikanyo et al., 2023).

Wearable health monitoring devices, such as ECG patches, smartwatches, and fitness trackers, continuously capture cardiovascular parameters including heart rate, rhythm, and blood pressure for real-time analysis. These monitors stream data to mobile or web-based applications that process the information using Artificial intelligence. Transformer-based models have demonstrated superior capability in capturing temporal dependencies within sequential health data (Xu et al., 2023). Their integration also allows early anomaly detection, individual feedback, as well as emergency notification.

mHealth applications facilitate remote patient monitoring by tracking symptoms, analyzing trends, and generating alerts when readings exceed normal thresholds. In clinical environments, real-time dashboards process multimodal patient data aggregated from diverse sources to support doctors' clinical decision-making (Kumar et al., 2023). Combination with Electronic Health Records (EHRs) improves monitoring and decision-making.

Notwithstanding these benefits, real-time CVD monitoring can be at risk of several problems; that is, sensor data are often altered by motion artifacts and environmental noise, which lowers accuracy (Khan et al., 2025). Implementing complex models like transformers on edge devices entails optimization to exceed latency and energy constraints. Additionally, real-time data transmission raises privacy and security worries, particularly in cloud-based platforms (Wang et al., 2025).

Another significant challenge is the interpretability of these AI models. Medical professionals often distrust black-box systems, particularly in high-stakes diagnoses. The absence of clinical validation through large-scale randomized controlled trials continues to impede regulatory approval and broader implementation. Furthermore, infrastructural limitations in low- and middle-income countries

constrain access to these technologies, despite the disproportionately high burden of CVDs (Sokol et al., 2025).

G. Future Directions

Future work on transformer-based multimodal IoT data fusion for real-time smart healthcare should focus more on developing light architectures for easy edge deployment. These models need to maintain diagnostic accuracy at lower computational cost and enable real-time CVD monitoring on wearable devices and low-power sensors (Ahmad et al., 2025; Khan et al., 2025). At the same time, federated learning techniques can support privacy-preserving collaboration across decentralized healthcare nodes without transferring sensitive patient data, aligning with privacy laws and ethical standards (Shukla et al., 2025). Unified multimodal pretraining represents a promising alternative, facilitating the learning of cross-modal relationships from ECG, PPG, imaging, and clinical notes to improve model robustness and generalization.

Transformer models also need to be adaptable to produce real-time predictive alerts by processing continuous streams of data and the estimation of uncertainty to avoid important events (e. g. heart attack) into clinical decision support systems (CDSS). They require interpretable outputs and adequate consistency with electronic health records to ensure clinician trust and usability (Shaik et al., 2024). Equality-focused benchmarking and policy implementation in resource-limited settings are critical for ensuring equitable global deployment of transformer-based healthcare innovations. Infrastructural limitations, gender-related disparities, and implementation challenges continue to hinder the effective integration of transformer models in clinical settings.

III. CONCLUSION

Transformer-based multimodal fusion is a paradigm shift in CVD monitoring with the integration of IoMT signal streams (ECG, PPG, imaging, EHRs) into unified diagnostic insights. By integrating physiological, clinical, and imaging data, these approaches offer a more holistic basis for prediction, which can translate to timelier, proactive patient management, especially in time-sensitive care settings. This review show that transformer-based models outperform traditional models on many data fusion tasks, yet the widespread application of transformer-based models in clinical settings is consistently hindered by numerous significant challenges: high computational complexity, poor interpretability, data heterogeneity, and poor deployment performance in resource-constrained settings.

In addition, the review points out the need to develop lightweight variants of transformers for inference on the device, merge explainable AI techniques for clinical trust considerations, and use learner paradigms that conserve privacy, such as federated learning. The deployment of real-time systems remains a developing challenge, requiring careful optimization of latency, energy efficiency, and predictive accuracy. Moreover, the absence of standardized benchmarks for multimodal CVD applications impedes cross-study comparisons and limits the generalizability of findings. Integrating streams of IoT data and simply applying transformer-based models for real-time monitoring and prediction cannot be achieved until all relevant stakeholders in the health sector are brought together to turn the theoretical approach into something that happens in real life.

Achieving equitable global health advancements necessitates overcoming significant infrastructural, ethical, and socioeconomic barriers. This is particularly critical in developing nations such as Nigeria, where constraints in essential infrastructure, funding, trained personnel, and technology persist.

Author Contribution

Author 1 Contributions

Author 1 was primarily responsible for the introductory research and theoretical analysis of the review. His specific contributions include:

- Literature review on Transformer-based architectures (e.g., TabTransformer, BERT, ViT) for multimodal data fusion.
- Performing the critical evaluation of existing models for cardiovascular disease monitoring tasks like coronary heart disease and stroke prediction.
- Drafting the core manuscript sections on fusion methodologies and the comparative analysis of model performance.

Author 2 Contributions

Author 2 led the analysis of practical deployment challenges and the forward-looking perspective of the review. His specific contributions include:

- Conducting the investigation and categorization of multimodal IoT data in healthcare, such as ECG, PPG, and EHR.
- Critically examining the constraints and optimization techniques for deploying Transformers in resource-constrained IoT and edge environments.
- Synthesizing strategies for model compression, including knowledge distillation and quantization, for real-time monitoring.
- Developing the framework for real-time deployment architectures and identifying key research gaps to formulate future directions for scalable AI in healthcare.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval

There is no ethical issue.

REFERENCES:

[1] Abdar, M., Salari, S., Qahremani, S., Lam, H. K., Karray, F., Hussain, S., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2023). UncertaintyFuseNet: Robust uncertainty-aware hierarchical feature fusion model with Ensemble Monte Carlo Dropout for COVID-19 detection. *Information Fusion*, *90*, 364-381. https://doi.org/10.1016/j.inffus.2022.09.023

[2] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2017). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine, 100,* 270–278. https://doi.org/10.1016/j.compbiomed.2017.09.017

[3] Adedinsewo, D., Hardway, H. D., Morales-Lara, A. C., Wieczorek, M. A., Johnson, P. W., Douglass, E. J., Dangott, B. J., Nakhleh, R. E., Narula, T., Patel, P. C., Goswami, R. M., Lyle, M. A., Heckman, A. J., Leoni-Moreno, J. C., Steidley, D. E., Arsanjani, R., Hardaway, B., Abbas, M., Behfar, A., Attia, Z. I., … Yamani, M. (2023). Non-invasive detection of cardiac allograft rejection among heart transplant recipients using an electrocardiogram based deep learning model. *European heart journal. Digital health*, *4*(2), 71–80. https://doi.org/10.1093/ehjdh/ztad001

[4] Adeyi, V., Xiaoling, Z., Uko, K., Kanjarawy, I., & Emmanuel, G. (2025). CardioNet+: Revolutionizing Heart Failure Diagnosis with Multi-modal Learning. *Available at SSRN 5132486*.

[5] Ahmad, M., Ahmed, A., Hashim, H., Farsi, M., and Mahmoud, N. (2025). Enhancing heart disease diagnosis using ECG signal reconstruction and deep transfer learning classification with optional SVM integration. *Diagnostics, 15*(12), 1501. https://doi.org/10.3390/diagnostics15121501

[6] Alam, F., Ananbeh, O., Malik, K. M., Odayani, A. A., Hussain, I. B., Kaabia, N., ... & Saudagar, A. K. J. (2023). Towards predicting length of stay and identification of cohort risk factors using self-attention-based transformers and association mining: COVID-19 as a phenotype. *Diagnostics*, *13*(10), 1760.

[7] Alimbayeva, Z., Alimbayev, C., Ozhikenov, K., Bayanbay, N., and Ozhikenova, A. (2024). *Wearable ECG device and machine learning for heart monitoring*. *Sensors*, *24*(13), 4201. https://doi.org/10.3390/s24134201

[8] AL-Mosawi, R. H., & Al-Shammari, A. (2024). A Survey on Exploring Multimodal Fusion in Healthcare: Challenges and Solutions. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *16*(4), 318-328.

[9] Al-Shareeda, M., Alsadhan, A., Qasim, H., & Manickam, S. (2023). Long range technology for internet of things: Review, challenges, and future directions. *Bulletin of Electrical Engineering and Informatics, 12*(6), 3758–3767. https://doi.org/10.11591/eei.v12i6.5214

[10] Anand, D., Singhal, V., Shanbhag, D. D., KS, S., Patil, U., Bhushan, C., ... & Kass-Hout, T. (2023). One-shot localization and segmentation of medical images with foundation models. *arXiv preprint arXiv:2310.18642*.

[11] Ansari, Y., Mourad, O., Qaraqe, K., & Serpedin, E. (2023). Deep learning for ECG arrhythmia detection and classification: An overview of progress for period 2017–2023. *Frontiers in Physiology, 14*, 1246746. https://doi.org/10.3389/fphys.2023.1246746

[12] Assen, M. V., Tariq, A., Razavi, A. C., Yang, C., Banerjee, I., & De Cecco, C. N. (2023). *Fusion Modeling: Combining Clinical and Imaging Data to Advance Cardiac Care*. *Circulation: Cardiovascular Imaging*, 16:e014533. https://doi.org/10.1161/CIRCIMAGING.122.014533.

[13] Banimfreg, B. H. (2023). A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics. *Healthcare Analytics, 3*, 100190. https://doi.org/10.1016/j.health.2023.100190

[14] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., ... & Oktay, O. (2023). Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15016-15027).

[15] Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., & Xu, Y. (2024). *Autoencoders and their applications in machine learning: a survey*. Artificial Intelligence Review, 57:28. https://doi.org/10.1007/s10462-023-10662-6

[16] Berikol, G. B., Kanbakan, A., Ilhan, B., & Doğanay, F. (2025). Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education. *Turkish Journal of Emergency Medicine, 25*(2), 67–91. DOI: 10.4103/tjem.tjem_45_25

[17] Boikanyo, K., Zungeru, A. M., Sigweni, B., Yahya, A., & Lebekwe, C. (2023). Remote patient monitoring systems: Applications, architecture, and challenges. *Scientific African, 20*, e01638. https://doi.org/10.1016/j.sciaf.2023.e01638

[18] Bonanno, M., Saracino, B., Ciancarelli, I., Panza, G., Manuli, A., Morone, G., & Calabrò, R. S. (2025). Assistive Technologies for Individuals with a Disability from a Neurological Condition: A Narrative Review on the Multimodal Integration. *Healthcare*, *13*(13), 1580. https://doi.org/10.3390/healthcare13131580

[19] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision* (pp. 205-218). Cham: Springer Nature Switzerland.

[20] Ceccarelli, M., Taje, R., Papuc, P. E., & Ambrogi, V. (2022). An Analysis of Respiration with the Smart Sensor SENSIRIB in Patients Undergoing Thoracic Surgery. *Sensors*, *22*(4), 1561. https://doi.org/10.3390/s22041561

[21] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). *TransUNet: Transformers make strong encoders for medical image segmentation*. arXiv. https://doi.org/10.48550/arXiv.2102.04306

[22] Chen, X., Xie, H., Tao, X., Wang, F. L., Leng, M., & Lei, B. (2024). Artificial intelligence and multimodal data fusion for smart healthcare: Topic modeling and bibliometrics. *Artificial Intelligence Review, 57*, 91. https://doi.org/10.1007/s10462-024-10712-7

[23] Chen X, Liao S, Wang P, Huang G, Li J, et al. (2025) Multimodal Data Fusion-Based Risk Assessment Models and Clinical Decision Support Systems for Intraoperative Acquired Pressure Injuries. J Surg 10: 11261 https://doi.org/10.29011/2575-9760.011261

[24] Creţu, I., Tindale, A., Abbod, M., Khir, A., Balachandran, W., and Meng, H. (2023). Multimodal arrhythmia classification using deep neural networks. *Proceedings of the International Conference on Biomedical Engineering and Systems (ICBES 2023).* https://doi.org/10.11159/icbes23.152

[25] Di Martino, F., & Delmastro, F. (2025). *Challenges and limitations in the synthetic generation of mHealth sensor data*. arXiv. https://arxiv.org/abs/2505.14206

[26] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[27] Farrokhi, M., Ranjbar, Z., Taheri, F., Moeini, A., & Farahmand, S.M. (2024). *Artificial Intelligence for Remote Patient Monitoring: Advancements, Applications, and Challenges*. Kindle and PreferPub. https://www.researchgate.net/publication/379775995

[28] Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2025). A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing, 158*, 105509. https://doi.org/10.1016/j.imavis.2025.105509

[29] Gupta, P., Verma, D., & Gupta, A. (2024). Unveiling the layered architecture of IoT: A comprehensive overview. In *Handbook of Research on Emerging Trends and Applications of the Internet of Things* (pp. 141–163). IGI Global. https://doi.org/10.4018/979-8-3693-2373-1.ch008

[30] Gutheil, J., & Donsa, K. (2022). SAINTENS: Self-Attention and Intersample Attention Transformer for Digital Biomarker Development Using Tabular Healthcare Real World Data. *Studies in health technology and informatics*, *293*, 212–220. https://doi.org/10.3233/SHTI220371

[31] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574-584).

[32] Henke, C. (2022, November 3). *A comprehensive guide to IoT protocols*. Accessed 15th July 2025 from https://www.emnify.com/blog/guide-iot-protocols

[33] Islam, M. N., Raiyan, K. R., Mitra, S., Mannan, M. M. R., Tasnim, T., Putul, A. O., & Mandol, A. B. (2023). Predictis: An IoT and machine learning-based system to predict risk level of cardio-vascular diseases. *BMC Health Services Research, 23*(171). https://doi.org/10.1186/s12913-023-09104-4

[34] Jia, H., Zhang, J., Ma, K., Qiao, X., Ren, L., & Shi, X. (2024). Application of convolutional neural networks in medical images: A bibliometric analysis. *Quantitative Imaging in Medicine and Surgery, 14*(5), 3501–3518. https://doi.org/10.21037/qims-23-1600

[35] John, A., Cardiff, B., and John, D. (2024). *A review on multisensor data fusion for wearable health monitoring*. arXiv. https://arxiv.org/abs/2412.05895v1

[36] Kasartzian, D., & Tsiampalis, T. **(2025).** Transforming cardiovascular risk prediction: A review of machine learning and artificial intelligence innovations. *Life, 15*, 94. https://doi.org/10.3390/life15010094

[37] Kern, J., Quintero Bernal, D. F., & Urrea, C. (2025). *Multimodal data fusion system for* accurate identification of impact points on rocks in mining comminution tasks. *Processes*, 13(1), 87. https://doi.org/10.3390/pr13010087

[38] Khan, B., Khan, W., Masrur, M. H., Khalid, R. T., Awais, M., Khan, B., Khoo, B. L., & Abdullah, S. (2025). Hybrid sensor integration in wearable devices for improved cardiovascular health monitoring. *Journal of Science: Advanced Materials and Devices, 10*(2), 100889. https://doi.org/10.1016/j.jsamd.2025.100889

[39] Kim, K. B., and Baek, H. J. (2023). Photoplethysmography in wearable devices: A comprehensive review of technological advances, current challenges, and future

directions. *Electronics, 12*(13), 2923. https://doi.org/10.3390/electronics12132923

[40] Koulouras, G., Katsoulis, S., & Zantalis, F. (2025). Evolution of Bluetooth Technology: BLE in the IoT Ecosystem. *Sensors*, *25*(4), 996. https://doi.org/10.3390/s25040996

[41] Krones, F., Marikkar, U., Parsons, G., Szmul, A., & Mahdi, A. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion, 114*, 102690. https://doi.org/10.1016/j.inffus.2024.102690

[42] Kulasekara, M., Inglés-Romero, J. F., Imbernón, B., and Abellan, J.L. (2025). Saffe: Multimodal model composition with semantic-alignment fusion of frozen encoders. *Journal of Supercomputing, 81*, 1114. https://doi.org/10.1007/s11227-025-07473-7

[43] Kumar, D., Hasan, Y., & Afroz, S. (2023). Mobile health monitoring system: A comprehensive review. *International Journal of Research Publication and Reviews, 4*, 1922–1954. https://doi.org/10.55248/gengpi.4.623.45128

[44] Kumar, S., Rani, S., Sharma, S., & Min, H. (2024). Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review. *Bioengineering (Basel, Switzerland)*, *11*(12), 1233. https://doi.org/10.3390/bioengineering11121233

[45] Lentzen, M., Linden, T., Veeranki, S., Madan, S., Kramer, D., Leodolter, W., & Fröhlich, H. **(2023).** A transformer-based model trained on large scale claims data for prediction of severe COVID-19 disease progression. *IEEE Journal of Biomedical and Health Informatics, 27*(9), 4548–4558.

[46] Li, C., Wang, J., Wang, S., & Zhang, Y. (2024). A review of IoT applications in healthcare. *Neurocomputing*, *565*, 127017.

[47] Morid, M. A., Sheng, O. R. L., & Dunbar, J. (2023). Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems, 14*(1), Article 2. https://doi.org/10.1145/3531326

[48] Li, J., & Zhang, Y. (2024). Regressive vision transformer for dog cardiomegaly assessment. *Scientific Reports*, *14*(1), 1539.

[49] Li, S., and Tang, H. (2024). *Multimodal alignment and fusion: A survey*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Advance online publication. https://arxiv.org/abs/2411.17040v1

[50] Lu, H., Feng, X. and Zhang, J. (2024). Early detection of cardiorespiratory complications and training monitoring using wearable ECG sensors and CNN. *BMC Med Inform Decis Mak* **24**, 194. https://doi.org/10.1186/s12911-024-02599-9

[51] Madan, S., Lentzen, M., Brandt, J., et al. (2024). Transformer models in biomedicine. *BMC Medical Informatics and Decision Making, 24*, 214. https://doi.org/10.1186/s12911-024-02600-5

[52] Mansouri, S., Alharbi, Y., Alshrouf, A., & Alqahtani, A. (2022). Cardiovascular diseases diagnosis by impedance cardiography. *Journal of Electrical Bioimpedance, 13*, 88–95. https://doi.org/10.2478/joeb-2022-0013

[53] Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine, 157*, 106791. https://doi.org/10.1016/j.compbiomed.2023.106791

[54] Moon, K. S., & Lee, S. Q. (2023). A Wearable Multimodal Wireless Sensing System for Respiratory Monitoring and Analysis. *Sensors*, *23*(15), 6790. https://doi.org/10.3390/s23156790

[55] Nejad, S. M. (2025). Hybrid of convolutional neural network and support vector machine for cancer type prediction. *Control and Optimization in Applied Mathematics, 10*(1), 73–89. https://doi.org/10.30473/coam.2025.72710.12

[56] Noor, N., Bilal, M., Abbasi, S.F., Pournik, O. and Arvanitis, T.N. (2025) A novel transformer-based approach for cardiovascular disease

detection. Front. Digit. Health 7:1548448. doi: 10.3389/fdgth.2025.1548448

[57] Osama, M., Ateya, A. A., Sayed, M. S., Hammad, M., Pławiak, P., Abd El-Latif, A. A., & Elsayed, R. A. (2023*). Internet of medical things and Healthcare 4.0: Trends, requirements, challenges, and research directions. Sensors, 23*(17), 7435. https://doi.org/10.3390/s23177435

[58] Paul, S. G., Saha, A., Hasan, M. Z., Noori, S. R. H., & Moustafa, A. (2024). A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions. *IEEE Access*. Advance online publication. https://doi.org/10.1109/ACCESS.2024.3354809

[59] Psomakelis, E., Makris, A., Tserpes, K., & Pateraki, M. (2023). A lightweight storage framework for edge computing infrastructures: EdgePersist. *Software Impacts, 17*, 100549. https://doi.org/10.1016/j.simpa.2023.100549

[60] Puttagunta, M., Subban, R., & Babu, N. K. (2022)**.** SwinIR Transformer applied for medical image super-resolution. *Procedia Computer Science, 204*, 907–913. https://doi.org/10.1016/j.procs.2022.08.110

[61] Raj, R., & Kos, A. (2024). Study of Human–Robot Interactions for Assistive Robots Using Machine Learning and Sensor Fusion Technologies. *Electronics*, *13*(16), 3285. https://doi.org/10.3390/electronics13163285

[62] **R**asmy, L., Xiang, Y., Xie, Z., et al. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine, 4*, 86. https://doi.org/10.1038/s41746-021-00455-y

[63] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., Criqui, M., DeCleene, N., Eagle, K. A., Emmons-Bell, S., Feigin, V. L., Fernández-Solà, J., Fowkes, G., Gakidou, E., Grundy, S. M., He, F. J., Howard, G., Hu, F., Inker, L., Karthikeyan, G., Kassebaum, N., Koroshetz, W., Lavie, C., Lloyd-Jones, D., Lu, H. S., Mirijello, A., Misganaw, A. T., Mokdad, A., Moran, A. E., Muntner, P., Narula, J., Neal, B., Ntsekhe, M., de Oliveira, G. M. M., Otto, C., Owolabi, M. O., Pratt, M., Rajagopalan, S., Reitsma, M. B., Ribeiro, A. L. P., Rigotti, N. A., Rodgers, A., Sable, C. A., Shakil, S. S., Sliwa, K., Stark, B. A., Sundström, J., Timpel, P., Tleyjeh, I. I., Valgimigli, M., Vos, T., Whelton, P. K., Yacoub, M., Zuhlke, L. J., Abbasi-Kangevari, M., Abdi, A., Abedi, A., Aboyans, V., Abrha, W. A., Abu-Gharbieh, E., Abushouk, A. I., Acharya, D., Adair, T., Adebayo, O. M., Ademi, Z., Advani, S. M., Afshari, K., Afshin, A., Agarwal, G., Agasthi, P., Ahmad, S., Ahmadi, S., Ahmed, M. B., Aji, B., Akalu, Y., Akande-Sholabi, W., Aklilu, A., Akunna, C. J., Alahdab, F., Al-Eyadhy, A., Alhabib, K. F., Alif, S. M., Alipour, V., Aljunid, S. M., Alla, F., Almasi-Hashiani, A., Almustanyir, S., Al-Raddadi, R. M., Amegah, A. K., Amini, S., Aminorroaya, A., Amu, H., Amugsi, D. A., Ancuceanu, R., Anderlini, D., Andrei, T., Andrei, C. L., Ansari-Moghaddam, A., Anteneh, Z. A., Antonazzo, I. C., Antony, B., Anwer, R., Appiah, L. T., Arabloo, J., Ärnlöv, J., Artanti, K. D., Ataro, Z., Ausloos, M., Avila-Burgos, L., Awan, A. T., Awoke, M. A., Ayele, H. T., Ayza, M. A., Azari, S., ... Murray, C. J. L., & Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. *Journal of the American College of Cardiology, 76*(25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010

[64] Sagheer, S.V.M., K H, M., Ameer, P.M., Parayangat, M., Abbas, M. (2025). Transformers for Multi-Modal Image Analysis in Healthcare. *Computers, Materials & Continua*, *84(3)*, 4259–4297. https://doi.org/10.32604/cmc.2025.063726

[65] Saleh, H., McCann, M., El-Sappagh, S., & Breslin, J. G. (2025). TransformerFusionNet: A Real-Time Multimodal Framework for ICU Heart Failure Mortality Prediction Using Big Data Streaming.

[66] Shaik, T., Tao, X., Li, L., Xie, H., & Velásquez, J. D. (2024). A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion, 102*, 102040. https://doi.org/10.1016/j.inffus.2023.102040

[67] Shukla, S., Rajkumar, S., Sinha, A., Esha, M., Elango, K. and Sampath, V. (2025). Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity. *Scientific Reports, 15*, 13061. https://doi.org/10.1038/s41598-025-95858-2

[68] Sokol, K., Fackler, J., & Vogt, J. E. (2025). Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. *NPJ digital medicine*, *8*(1), 345. https://doi.org/10.1038/s41746-025-01725-9

[69] Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Pinho Gomes, A. C., Payberah, A. H., Zottoli, M., Nazarzadeh, M., Conrad, N., Rahimi, K., & Salimi-Khorshidi, G. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics, 101*, 103337. https://doi.org/10.1016/j.jbi.2019.103337

[70] Sudha, V. K., & Kumar, D. (2023). Hybrid CNN and LSTM network for heart disease prediction. *SN Computer Science, 4*. https://doi.org/10.1007/s42979-022-01598-9

[71] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24* (pp. 36-46). Springer International Publishing.

[72] Valeria, C., Cristina, V., Giulia, D. V., Chiara, M., Valentina, C., & Giampaolo, N. (2024). Psychological risk factors and cardiovascular disease. *Frontiers in Psychology, 15*. https://doi.org/10.3389/fpsyg.2024.1419731

[73] Wang, Y. R., Yang, K., Wen, Y., et al. (2024). Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. *Nature Medicine, 30*, 1471–1480. https://doi.org/10.1038/s41591-024-02971-2

[74] Wang, W.H., Hsu, W.S. (2023). Integrating Artificial Intelligence and Wearable IoT System in Long-Term Care Environments. *Sensors*. 23(13):5913. https://doi.org/10.3390/s23135913

[75] Wang, X., Xu. Z. and Sui, X. (2025). Intelligent data analysis in edge computing with large language models: applications, challenges, and future directions. Front. Comput. Sci. 7:1538277. doi: 10.3389/fcomp.2025.1538277

[76] World Health Organization. (2021, June 11). *Cardiovascular diseases (CVDs)* (Fact sheet). Accessed 14th June 2025 from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[77] Xing, Y., Yang, Y., Yang, K., Lu, A., Xing, L., Mackie, K., & Guo, F. (2025). Intelligent sensing devices and systems for personalized mental health. *Med-X*, *3*(1), 10.

[78] Xu, P., Zhu, X., & Clifton, D. A. (2023). *Multimodal learning with transformers: A survey. arXiv.* https://arxiv.org/abs/2206.06488

[79] Yang, F., Qiao, Y., Hajek, P., & Abedin, M. Z. (2024). Enhancing cardiovascular risk assessment with advanced data balancing and domain knowledge-driven explainability. *Expert Systems with Applications, 255*(Part D), 124886. https://doi.org/10.1016/j.eswa.2024.124886

[80] Yan, J., Tian, J., Yang, H., Han, G., Liu, Y., He, H., Han, Q., & Zhang, Y. (2022). *A clinical decision support system for predicting coronary artery stenosis in patients with suspected coronary heart disease.* Computers in Biology and Medicine, *151*, 106300. https://doi.org/10.1016/j.compbiomed.2022.106300

[81] Yıldırım, F., Yalman, Y., Bayındır, K. Ç., & Terciyanlı, E. (2025). Comprehensive Review of Edge Computing for Power Systems: State of the Art, Architecture, and Applications. *Applied Sciences*, *15*(8), 4592. https://doi.org/10.3390/app15084592

[82] Zhong, M., Van Zoest, V., Bilal, A. M., Papadopoulos, F. C., & Castellano, G. (2022). Unimodal vs. multimodal prediction of antenatal depression from smartphone-based survey data in a longitudinal study. *Proceedings of the 24th ACM International Conference on Multimodal Interaction (ICMI '22),* November 7–11, 2022, Bengaluru, India. ACM. https://doi.org/10.1145/3536221.3556605

[83] Zhou, H. Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., ... & Li, W. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, *7*(6), 743-755.

[84] Zhu, G., Deng, E., Qin, Z., Khan, F., Wei, W., Srivastava, G., Xiong, H., & Kumari, S. (2024)**.** Cross-modal interaction and multi-source visual fusion for video generation in fetal cardiac screening. *Information Fusion, 111*, 102510. https://doi.org/10.1016/j.inffus.2024.102510