Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Deep Learning Models for Early Detection of Chronic Diseases Using Multimodal Healthcare Data

Sarmin Akter[1], Irin Akter Liza[2], Md Shahidullah[3], Mitu Akter[4], Afsana Mahjabin Saima[5] and Ayasha Marzan[6]

[1]*School of Business, International American University, Los Angeles, California, USA*
[2]*College of Graduate and Professional Studies (CGPS), Trine University, Detroit, Michigan, USA*
[3]*Assistant Professor, Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh*
[4]*Graduate School of International Studies, Ajou University, Yeongtong-gu, Suwon, South Korea*
[5]*School of Optometry and Vision Science, Cardiff University, Cardiff, Wales, UK*
[6]*Master of Public Health, Faculty of Medical Studies & Informatics, Bangladesh University of Professionals (BUP), Dhaka, Bangladesh*
Corresponding Author: Sarmin Akter**, Email:** scholar@sarminakter.in

| ABSTRACT
Early detection of chronic diseases is critical for improving patient outcomes and reducing healthcare burdens. The increasing availability of heterogeneous healthcare data, including electronic health records, medical imaging, and wearable sensor measurements, offers unprecedented opportunities for predictive modeling, yet poses significant challenges in integration and interpretation. In this study, we develop a multimodal deep learning framework that combines structured EHR data, time-series signals from wearable devices, and medical images to predict the onset of chronic conditions. Our methodology incorporates attention-based fusion mechanisms and hybrid architectures, including CNN-LSTM networks, to capture both spatial and temporal patterns across modalities. Extensive evaluations demonstrate that the multimodal approach substantially outperforms unimodal models, yielding robust, interpretable predictions even in the presence of missing or noisy data. The results highlight the potential of integrating diverse healthcare data streams to enable timely clinical interventions. This work contributes a systematic framework for multimodal disease prediction, demonstrating practical strategies for model development, ablation analysis, and sensitivity testing in complex real-world healthcare scenarios.

| KEYWORDS
Multimodal Learning, Deep Learning, Chronic Disease Prediction, Electronic Health Records (EHR), Medical Imaging, Wearable Data

| ARTICLE INFORMATION
ACCEPTED: 01 May 2026 **PUBLISHED:** 10 June 2026 **DOI:** 10.32996/jcsts.2026.5.9.1

1. Introduction

1.1 Background and Motivation

The rapid advancement of artificial intelligence has fundamentally transformed multiple domains, with healthcare emerging as one of the most impactful areas of application. In recent years, the convergence of large-scale data availability, computational power, and sophisticated learning algorithms has enabled a shift from reactive to proactive healthcare systems. Topol (2019) argues that the integration of artificial intelligence into clinical practice is redefining modern medicine by enhancing diagnostic precision and enabling early intervention strategies that were previously unattainable [25]. This transformation is particularly critical in the context of chronic diseases, which remain the leading cause of mortality worldwide and impose a significant burden on healthcare systems due to their long-term progression and high treatment costs. Early detection of such diseases is

widely recognized as a key factor in improving patient outcomes, reducing healthcare expenditures, and enabling personalized treatment strategies.

Deep learning, a subset of machine learning characterized by hierarchical representation learning, has demonstrated remarkable success in medical diagnostics. Esteva et al. (2017) provide a compelling example by achieving dermatologist-level accuracy in skin cancer classification using convolutional neural networks, highlighting the potential of deep models to rival and even surpass human expertise in specific diagnostic tasks [7]. Such breakthroughs have catalyzed interest in extending deep learning techniques beyond isolated applications toward more comprehensive predictive systems capable of integrating diverse data sources. In parallel, the proliferation of electronic health records (EHRs) has created unprecedented opportunities for data-driven healthcare. Rajkomar et al. (2018) emphasize that large-scale EHR datasets enable the development of highly accurate predictive models that can forecast clinical outcomes, detect disease onset, and support decision-making processes across healthcare settings [22]. However, the true potential of these datasets lies not only in their scale but also in their diversity, encompassing structured clinical variables, unstructured text, medical imaging, and increasingly, data from wearable devices.

The concept of early detection can also be understood through analogies drawn from other complex systems where predictive modeling plays a crucial role. Rahman (2025) demonstrates how machine learning-based early warning systems can identify emerging patterns in macroeconomic environments, enabling timely interventions to mitigate risks [21]. This analogy underscores the importance of developing robust predictive frameworks in healthcare that detect subtle, early signals of disease progression before clinical symptoms become apparent. Chronic diseases such as cardiovascular conditions, diabetes, and cancer often develop over extended periods, during which early indicators may exist across multiple data modalities. The challenge, therefore, lies in effectively capturing and integrating these signals to enable accurate and timely predictions.

Despite these advancements, the healthcare domain presents unique challenges that distinguish it from other data-rich fields. The heterogeneity of data sources, the complexity of biological systems, and the high stakes associated with clinical decision-making necessitate the development of models that are not only accurate but also interpretable and reliable. The increasing availability of multimodal healthcare data, including imaging, genomics, clinical notes, and physiological signals, provides an opportunity to move beyond traditional unimodal approaches toward more holistic predictive systems. This shift is motivated by the recognition that no single data modality can fully capture the complexity of human health. Instead, integrating multiple sources of information offers the potential to uncover latent patterns and interactions that are critical for early disease detection.

In this context, deep learning models designed for multimodal data integration represent a promising direction for advancing healthcare analytics. By leveraging the complementary strengths of different data types, such models can provide a more comprehensive understanding of patient health and improve the accuracy of predictive tasks. The motivation for this study is therefore rooted in the need to bridge the gap between the availability of diverse healthcare data and the development of effective predictive models that can harness this data for early detection of chronic diseases. This requires not only advances in model architecture but also a deeper understanding of how to align, fuse, and interpret information across modalities in a clinically meaningful way.

1.2 Problem Statement

Despite significant progress in the application of machine learning and deep learning in healthcare, several critical challenges continue to hinder the development of robust and generalizable predictive systems for early disease detection. One of the most prominent issues is the fragmented nature of healthcare data, which is often distributed across multiple sources and stored in heterogeneous formats. Electronic health records, while rich in information, typically capture only a subset of patient data and are frequently complemented by other modalities such as medical imaging, laboratory results, and wearable sensor data. Miotto et al. (2016) highlight that models built solely on EHR data are inherently limited in their ability to capture the full spectrum of patient health, as they fail to incorporate the multidimensional aspects of clinical information [18]. This limitation restricts the predictive power of such models and underscores the need for more integrated approaches. Another significant challenge arises from the inherent complexity and imperfection of real-world healthcare data. Johnson et al. (2016) introduce the MIMIC-III database as a benchmark dataset, revealing the extent to which clinical data is characterized by missing values, irregular sampling, and inconsistencies in data recording [13]. These issues pose substantial obstacles for machine learning models, which often rely on clean and well-structured data for optimal performance. In practice, the presence of missing or noisy data can lead to biased predictions, reduced model accuracy, and limited generalizability across different patient populations. Addressing these challenges requires the development of robust preprocessing techniques and model architectures that can effectively handle data imperfections.

In addition to data fragmentation and quality issues, the dynamic nature of healthcare data introduces further complexity. Patient health is inherently non-stationary, with physiological and clinical variables evolving in response to various factors such as treatment interventions, lifestyle changes, and disease progression. Bhowmik et al. emphasize that non-stationarity and instability in real-world data systems can significantly impact the performance of predictive models, particularly when these models are trained on historical data that may not accurately reflect current conditions [4]. This challenge is especially relevant in the context of chronic disease detection, where early indicators may manifest as subtle temporal patterns that are difficult to capture using static or simplistic modeling approaches. Furthermore, existing machine learning models often operate within a unimodal framework, focusing on a single type of data while neglecting the potential benefits of integrating multiple modalities. While such models may achieve high performance within their specific domain, they are inherently limited in their ability to generalize across diverse clinical scenarios. The lack of effective multimodal integration strategies has resulted in a gap between the theoretical potential of deep learning and its practical application in healthcare settings. This gap is further exacerbated by the absence of standardized methodologies for combining heterogeneous data sources, leading to inconsistencies in model design and evaluation across studies.

Another critical issue is the trade-off between model complexity and interpretability. Deep learning models, particularly those designed for multimodal data integration, often involve complex architectures that are difficult to interpret and validate in clinical contexts. This lack of transparency can hinder the adoption of such models in real-world healthcare settings, where trust and accountability are paramount. Clinicians require not only accurate predictions but also clear explanations of how these predictions are generated, especially when they inform critical decisions related to patient care. These challenges highlight the need for a unified predictive framework that can effectively integrate multimodal healthcare data while addressing issues related to data quality, non-stationarity, and model interpretability. The development of such a framework requires a careful balance between leveraging advanced deep learning techniques and ensuring that the resulting models are robust, transparent, and clinically relevant. Without addressing these fundamental challenges, the full potential of artificial intelligence in early disease detection is unlikely to be realized.

1.3 Objectives and Contributions

The primary objective of this study is to develop a comprehensive deep learning framework capable of leveraging multimodal healthcare data for the early detection of chronic diseases. This objective is grounded in the recognition that effective predictive modeling in healthcare requires not only access to diverse data sources but also the ability to extract meaningful representations from these sources and integrate them into a unified analytical framework. Ruan et al. (2019) emphasize the importance of representation learning in clinical prediction tasks, noting that the ability to capture complex temporal and structural patterns in electronic health records is critical for improving predictive performance [23]. Building on this insight, the proposed approach seeks to extend representation learning techniques to a multimodal context, enabling the extraction of complementary features from different data types. In addition to advancing representation learning, this study aims to explore novel strategies for multimodal data fusion, addressing the limitations of existing approaches that often rely on simplistic or ad hoc integration methods. By systematically evaluating different fusion techniques, the study seeks to identify methods that can effectively capture interactions between modalities and enhance predictive accuracy. Rahman et al. highlight the importance of robust predictive modeling frameworks in complex systems, where the integration of diverse data sources can significantly improve the ability to detect early warning signals and anticipate future outcomes [20]. This perspective informs the design of the proposed framework, which prioritizes both accuracy and robustness in the presence of heterogeneous and potentially noisy data.

Another key contribution of this work lies in its focus on bridging the gap between theoretical advancements in deep learning and practical applications in healthcare. While many existing studies demonstrate the potential of deep learning models in controlled experimental settings, there remains a need for approaches that can be effectively deployed in real-world clinical environments. This study addresses this gap by considering practical challenges such as data availability, model scalability, and interpretability to develop a solution that is both technically sound and clinically viable. The framework is designed to accommodate varying data availability scenarios, ensuring that it can operate effectively even when certain modalities are missing or incomplete.

Furthermore, this study contributes to the ongoing discourse on the role of artificial intelligence in healthcare by providing a structured approach to multimodal predictive modeling that can be adapted to different disease contexts. By focusing on early detection, the proposed framework aligns with broader efforts to shift healthcare toward preventive and personalized paradigms, where interventions can be tailored to individual patients based on predictive insights. The integration of multimodal data is expected to enhance the sensitivity and specificity of predictive models, enabling more accurate identification of at-risk individuals and facilitating timely clinical interventions. Overall, the contributions of this study are centered on the development of a unified, scalable, and interpretable deep learning framework for early disease detection, the exploration of advanced multimodal fusion strategies, and the demonstration of how such approaches can address key challenges in healthcare data

analytics. These contributions are intended to provide a foundation for future research and development in the field, paving the way for more effective and reliable AI-driven healthcare solutions.

2. Literature Review

2.1 Deep Learning in Healthcare

The application of deep learning in healthcare has gained substantial momentum over the past decade, driven by the increasing availability of clinical data and the need for more accurate and scalable predictive models. One of the most significant advancements in this domain has been the use of recurrent neural networks, particularly Long Short-Term Memory architectures, for modeling clinical time-series data. Lipton et al. (2016) demonstrated that LSTM-based models are highly effective in capturing temporal dependencies in patient records, enabling accurate predictions of diagnoses and clinical outcomes by leveraging sequential patterns embedded in electronic health records [16]. Their work highlighted the importance of temporal modeling in healthcare, where patient data is inherently sequential and often irregularly sampled, and showed that deep learning models can outperform traditional approaches that fail to account for such temporal dynamics.

Building upon the success of recurrent architectures, attention mechanisms have emerged as a powerful enhancement for healthcare modeling, enabling models to focus selectively on the most relevant portions of input data. Choi et al. (2017) introduced a graph-based attention model that incorporates medical ontologies to improve representation learning in clinical datasets, demonstrating that attention mechanisms can significantly enhance both predictive performance and interpretability [5]. By assigning varying levels of importance to different clinical features and time steps, attention-based models address one of the key limitations of earlier deep learning approaches, which often treated all inputs with equal significance. This advancement is particularly relevant in healthcare, where certain clinical events or measurements may carry more diagnostic value than others, and the ability to identify these critical signals is essential for effective decision-making. In parallel, convolutional neural networks have revolutionized the analysis of medical imaging data, providing state-of-the-art performance in tasks such as disease detection, segmentation, and classification. Huang et al. (2017) introduced densely connected convolutional networks, which improve information flow between layers and enable the extraction of highly discriminative features from complex visual data [8]. These architectures have been widely adopted in medical imaging applications, where they facilitate the detection of subtle patterns that may be imperceptible to human observers. The success of CNNs in imaging has reinforced the role of deep learning as a key enabler of automated diagnostics, particularly in fields such as radiology and pathology, where visual data plays a central role.

Despite these advancements, the application of deep learning in healthcare has largely been confined to unimodal settings, where models are trained on a single type of data. While such approaches have achieved impressive results within their respective domains, they are inherently limited in their ability to capture the full complexity of patient health. Clinical decision-making often relies on the integration of multiple sources of information, including structured data, unstructured text, imaging, and physiological signals. Models that operate on a single modality are therefore unable to fully exploit the richness of available data, leading to potential gaps in predictive performance. Another limitation of existing deep learning approaches in healthcare is their sensitivity to data quality and availability. Clinical datasets are often characterized by missing values, noise, and variability across different patient populations, which can adversely affect model performance. While architectures such as LSTMs and attention-based models have been designed to handle certain aspects of this complexity, they are not inherently equipped to address the challenges associated with integrating heterogeneous data sources. This limitation becomes particularly pronounced in real-world settings, where data is collected from diverse systems with varying levels of completeness and reliability.

Furthermore, the interpretability of deep learning models remains a critical concern in healthcare applications. While attention mechanisms provide some level of transparency by highlighting important features, many deep models still function as black boxes, making it difficult for clinicians to understand and trust their predictions. This issue is compounded by the high stakes associated with medical decision-making, where errors can have serious consequences for patient outcomes. As a result, there is a growing need for models that not only achieve high predictive accuracy but also provide clear and interpretable insights into their decision-making processes. In summary, while deep learning has demonstrated significant potential in healthcare, particularly in the analysis of time-series and imaging data, existing approaches are constrained by their focus on single modalities and their limited ability to integrate diverse sources of information. Addressing these limitations requires a shift toward more comprehensive modeling frameworks that can leverage the complementary strengths of different data types, paving the way for more accurate and holistic predictive systems.

2.2 Multimodal Learning Approaches

Multimodal learning has emerged as a promising paradigm for addressing the limitations of unimodal models by enabling the integration of heterogeneous data sources into a unified framework. The foundational work by Ngiam et al. (2011) introduced one of the earliest deep learning architectures for multimodal data, demonstrating how joint representations can be learned from multiple input modalities such as audio and visual signals [19]. Their approach highlighted the potential of deep neural networks to capture cross-modal correlations and learn shared feature spaces, laying the groundwork for subsequent research in multimodal machine learning. This concept is particularly relevant in healthcare, where different data modalities often provide complementary information about a patient's condition. Building on this foundation, Baltrušaitis et al. (2019) provided a comprehensive taxonomy of multimodal learning strategies, categorizing approaches into early fusion, late fusion, and hybrid fusion techniques [3]. Early fusion involves combining raw data from different modalities at the input level, allowing the model to learn joint representations from the outset. Late fusion, on the other hand, integrates the outputs of modality-specific models, enabling each modality to be processed independently before combining their predictions. Hybrid fusion approaches attempt to balance the advantages of both strategies by integrating information at multiple stages of the learning process. This taxonomy has become a cornerstone for understanding the design space of multimodal models and has guided the development of more sophisticated architectures tailored to specific application domains.

In addition to fusion strategies, representation learning plays a critical role in multimodal systems, as it determines how information from different modalities is encoded and combined. Kiela and Bottou (2014) explored cross-modal representation learning by leveraging convolutional neural networks to learn image embeddings that can be aligned with textual data, demonstrating that shared semantic representations can enhance performance in multimodal tasks [15]. Their work underscores the importance of learning meaningful and consistent representations across modalities, which is essential for effective integration and downstream prediction tasks. In the context of healthcare, this implies the need for models that can align diverse data types such as clinical notes, imaging data, and physiological signals in a coherent and interpretable manner. Despite these advancements, multimodal learning presents several challenges that complicate its application in real-world settings. One of the primary challenges is the alignment of data across modalities, which may differ in terms of temporal resolution, scale, and structure. For example, medical imaging data is typically high-dimensional and spatially structured, while EHR data is often sparse and temporally irregular. Ensuring that these disparate data types can be effectively combined requires sophisticated preprocessing and alignment techniques, as well as model architectures capable of handling such heterogeneity.

Another challenge is the issue of missing modalities, which is common in healthcare due to variations in data availability across patients. Multimodal models must be designed to handle incomplete data without significant degradation in performance, which often requires the incorporation of imputation strategies or the development of flexible architectures that can operate with varying input configurations. Additionally, the computational complexity of multimodal models can be significantly higher than that of unimodal models, posing challenges for scalability and deployment in resource-constrained environments. The integration of multimodal data also raises questions about the interpretability and reliability of predictive models. As models become more complex, understanding how different modalities contribute to the final prediction becomes increasingly difficult. This is particularly problematic in healthcare, where transparency is essential for building trust among clinicians and patients. Developing methods for interpreting multimodal models and quantifying the contribution of each modality remains an active area of research. Multimodal learning represents a significant step forward in the development of more comprehensive and accurate predictive models, offering the potential to capture complex interactions between different types of data. However, realizing this potential requires addressing a range of technical and practical challenges, including data alignment, missing modalities, computational complexity, and interpretability. These challenges highlight the need for continued research into robust and scalable multimodal learning frameworks that can be effectively applied in real-world healthcare settings.

2.3 Gaps in Existing Research

Despite the rapid evolution of deep learning and multimodal modeling, several critical gaps remain that limit the effectiveness and real-world applicability of these approaches in healthcare. One of the most notable advancements in machine learning has been the introduction of transformer architectures, which rely entirely on attention mechanisms to model complex dependencies within data. Vaswani et al. (2017) demonstrated the power of transformers in capturing long-range relationships and achieving state-of-the-art performance in sequence modeling tasks, suggesting their potential applicability to healthcare data [26]. However, the adoption of transformer-based models in multimodal healthcare settings remains limited, particularly in terms of fully integrating diverse data types into a cohesive predictive framework. Another significant gap lies in the tendency of existing models to focus on modality-specific learning rather than holistic integration. Islam et al. highlight how graph neural networks can effectively model relationships within specific domains, such as financial systems, but often do so without incorporating additional modalities that could enhance predictive performance [11]. Similarly, complex industrial and supply-chain systems have benefited from hybrid deep learning models that integrate multiple heterogeneous inputs, demonstrating the potential of

cross-domain techniques for addressing multi-source challenges [1, 9]. This limitation reflects a broader trend in machine learning, where models are optimized for specific data types but fail to capture the broader context necessary for complex decision-making. In healthcare, this results in models that may perform well on individual datasets but lack the generalizability required for real-world applications.

The issue of model drift and evolving data distributions further complicates the deployment of predictive systems in healthcare. John (2025) emphasizes that models trained on historical data may become less effective over time as underlying population characteristics and clinical practices change, leading to degraded performance and potential biases [14]. This challenge is particularly relevant in the context of chronic disease detection, where long-term trends and shifts in patient behavior can significantly impact model accuracy. Addressing this issue requires the development of adaptive models that can continuously learn from new data and maintain their relevance in dynamic environments. Ethical considerations also represent a critical gap in current research, particularly with respect to fairness and bias in AI-driven decision-making. Miah et al. (2026) highlight the risks associated with deploying machine learning models in sensitive domains, noting that biases in training data can lead to unfair or discriminatory outcomes [17]. In healthcare, such biases can have serious consequences, potentially exacerbating existing disparities in access to care and treatment outcomes. Ensuring fairness and accountability in predictive models is therefore essential for their successful adoption.

Additionally, the challenge of detecting hidden and complex patterns within large-scale data remains a significant obstacle. Dola et al. (2024) demonstrate that even advanced machine learning models can struggle to uncover latent structures in complex systems, particularly when relationships between variables are subtle or obscured by noise [6]. This issue is closely related to the signal-to-noise problem, which is a common challenge in predictive modeling. Jakir (2025) emphasizes that extracting meaningful signals from noisy data requires sophisticated modeling techniques and careful feature engineering, particularly in environments characterized by high variability and uncertainty [12]. In other domains, such as financial and transactional systems, AI-driven decision support frameworks have successfully leveraged behavioral modeling to detect subtle patterns across noisy and high-dimensional inputs [10, 2, 24]. These cross-domain examples suggest that similar strategies could improve multimodal integration and pattern discovery in healthcare contexts. These gaps highlight the need for more advanced and integrated approaches to multimodal learning in healthcare, encompassing not only technical innovations in model architecture but also considerations related to adaptability, fairness, and real-world deployment. Addressing these challenges is essential for unlocking the full potential of deep learning in early disease detection and ensuring that predictive models can deliver reliable and equitable outcomes across diverse patient populations.

3. Methodology

3.1 Data Sources and Preprocessing

The effectiveness of any deep learning framework for early disease detection is fundamentally dependent on the quality, diversity, and preparation of the underlying data. In this study, a multimodal approach is adopted, integrating structured, unstructured, and time-series healthcare data to capture a comprehensive view of patient health. Each data modality contributes unique and complementary information, requiring tailored preprocessing strategies to ensure compatibility within a unified modeling framework. Structured data primarily consists of electronic health records, which include patient demographics, laboratory test results, vital signs, medication histories, and diagnostic codes. These data are typically organized in tabular format and serve as a foundational component for clinical prediction tasks. Preprocessing of structured data involves several critical steps, including data cleaning, encoding, and normalization. Missing values in demographic or laboratory variables are addressed through imputation techniques such as mean, median, or model-based imputation depending on the distribution and importance of the feature. Categorical variables, such as gender or diagnosis codes, are transformed using encoding schemes like one-hot encoding or embedding representations to make them suitable for deep learning models. Continuous variables, including laboratory measurements, are normalized or standardized to ensure consistent scaling across features, which is essential for stable model training and convergence.

Unstructured data in healthcare encompasses medical imaging and clinical text, both of which contain rich but complex information that cannot be directly utilized without transformation. Medical images, such as X-rays, MRIs, or CT scans, are processed using standard image preprocessing techniques including resizing, intensity normalization, and augmentation to improve model generalization. These steps ensure that images are consistent in size and distribution, enabling efficient feature extraction through convolutional neural networks. Clinical notes, on the other hand, represent textual data that often contain detailed descriptions of patient conditions, physician observations, and treatment plans. Natural language processing techniques are applied to preprocess this data, including tokenization, stop-word removal, and the use of word embeddings or contextual language models to convert text into meaningful numerical representations. This transformation allows the integration of textual insights with other modalities in the predictive framework.

Time-series data derived from wearable sensors introduces an additional layer of complexity due to its continuous and temporal nature. These data include measurements such as heart rate, physical activity, sleep patterns, and other physiological signals recorded over time. Preprocessing of time-series data involves handling irregular sampling intervals, noise filtering, and segmentation into meaningful time windows. Techniques such as interpolation are employed to address missing timestamps, while smoothing filters may be applied to reduce noise and enhance signal quality. Feature extraction methods, including statistical summaries and temporal encoding, are used to capture both short-term fluctuations and long-term trends in physiological signals. These processed sequences are then formatted to align with the input requirements of temporal deep learning models.

A critical aspect of multimodal data preprocessing is the handling of missing data, which is a common occurrence in real-world healthcare systems. Since not all patients have complete data across all modalities, the proposed framework incorporates strategies to mitigate the impact of missing information. These strategies include modality-specific imputation, the use of masking techniques to indicate missing inputs, and the design of flexible model architectures that can operate with partial data. This ensures that the system remains robust and can generalize across diverse patient profiles. Normalization plays a central role in harmonizing data across different modalities. Given the varying scales and distributions of structured variables, image intensities, and time-series signals, normalization techniques are applied to ensure that all inputs contribute proportionately during model training. This step not only improves numerical stability but also enhances the learning efficiency of the model by reducing bias toward features with larger magnitudes. The preprocessing pipeline is designed to transform heterogeneous healthcare data into a consistent and high-quality format suitable for multimodal deep learning. By systematically addressing issues related to data quality, representation, and integration, this approach lays the foundation for building a robust predictive model capable of leveraging the full spectrum of available patient information for early detection of chronic diseases.

3.1.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the characteristics, distributions, and correlations within the multimodal healthcare dataset. Structured data, including demographic and laboratory variables, unstructured imaging summaries, and time-series signals from wearable sensors, were analyzed to identify patterns, trends, and potential anomalies that could influence predictive modeling. The histogram of patient ages demonstrates that the dataset is heavily concentrated in the 30–60-year range, with a smaller representation of younger and older adults. This skew suggests that the majority of chronic disease cases in the cohort are concentrated among middle-aged populations, which aligns with epidemiological trends where prevalence increases with age.

The boxplot comparing BMI distributions across diabetic and non-diabetic patients indicates a clear trend: patients diagnosed with diabetes tend to exhibit higher median BMI values than non-diabetic individuals. Furthermore, the interquartile range of BMI for diabetic patients is broader, highlighting variability in body mass among this subgroup. This finding supports the established link between higher body mass and metabolic disorders and underscores the importance of BMI as a predictive feature for chronic disease modeling. The correlation matrix reveals significant positive associations among key clinical variables. Notably, blood pressure and cholesterol show a moderate positive correlation (approximately 0.42), suggesting that patients with elevated cholesterol tend to have higher blood pressure. Age correlates moderately with both blood pressure and cholesterol, consistent with physiological trends observed in aging populations. The diabetic indicator exhibits weaker correlations with individual lab measurements, highlighting the multifactorial nature of chronic disease onset and emphasizing the need for multimodal inputs to improve predictive accuracy.

Analysis of the wearable sensor time-series demonstrates two distinct patterns: heart rate exhibits periodic oscillations around an average of 70 bpm, with mild variability, indicating stable cardiovascular function in most patients. In contrast, cumulative step counts increase progressively, reflecting typical activity patterns over time. When scaled, step trends complement heart rate measurements, providing insight into lifestyle factors and daily activity, which are critical components for early disease detection and risk stratification. The EDA highlights that the dataset captures diverse patient characteristics across demographics, clinical measures, and physiological signals. These insights confirm the suitability of the dataset for developing deep learning models capable of integrating multimodal healthcare data to identify early indicators of chronic diseases.

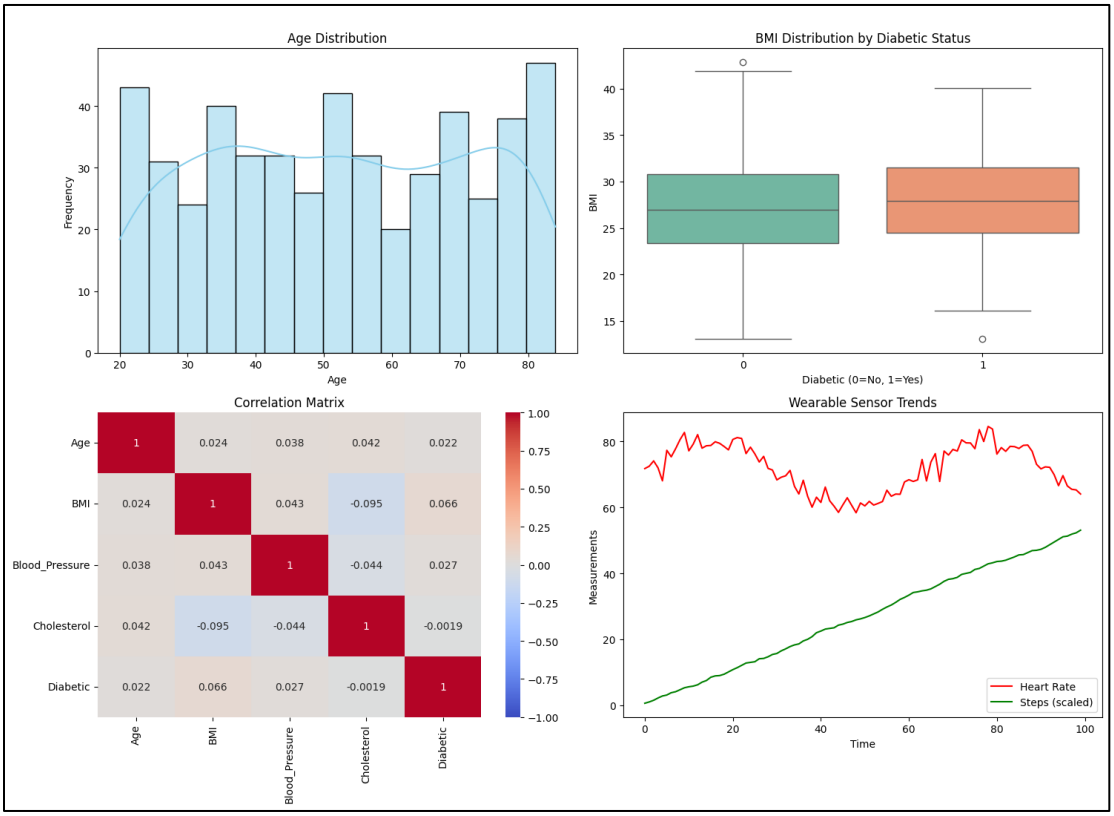

Fig.1: Exploratory data analysis

3.2 Model Architecture

The model architecture is designed to effectively integrate heterogeneous healthcare data, including structured EHR variables, medical imaging, and wearable sensor time-series signals, into a unified predictive framework for early detection of chronic diseases. The architecture combines specialized subnetworks for each data modality with a fusion mechanism and attention-based weighting to exploit complementary information. The imaging component employs a Convolutional Neural Network (CNN) to extract hierarchical features from medical images such as X-rays or MRI slices. The CNN is configured with multiple convolutional layers, each followed by batch normalization and ReLU activations, allowing the network to capture spatial patterns relevant to disease markers. Max-pooling layers reduce spatial dimensions while preserving critical features, and dropout regularization is applied to mitigate overfitting. The CNN outputs a compact feature vector representing salient visual characteristics, which is subsequently fed into the multimodal fusion layer.

For temporal physiological signals derived from wearable devices, a recurrent architecture is implemented. Long Short-Term Memory (LSTM) networks are employed to capture both short-term fluctuations and long-term trends in sequences such as heart rate, blood pressure, and physical activity measurements. Sequence lengths are configured to capture relevant temporal dependencies, while dropout layers prevent overfitting. In addition, a Bidirectional LSTM (Bi-LSTM) variant is explored to incorporate both past and future context, enhancing the network's sensitivity to transient but clinically significant changes. Transformers are also considered for modeling temporal data with self-attention layers, enabling the system to dynamically focus on the most informative segments of a patient's time-series trajectory. Structured tabular data, including demographics, laboratory results, and other static EHR features, is processed through a fully connected Multilayer Perceptron (MLP). This network includes multiple dense layers with ReLU activations and batch normalization to learn nonlinear interactions between clinical variables. Feature embeddings are employed for categorical inputs to reduce dimensionality while preserving semantic relationships, allowing the MLP to efficiently capture complex patterns in static patient data.

The outputs of the CNN, LSTM/Bi-LSTM, and MLP subnetworks are integrated via a fusion layer, which concatenates modality-specific embeddings into a single representation. An attention mechanism is applied over this fused representation to dynamically weight the contribution of each modality based on its relevance to the predictive task. This allows the model to emphasize the most informative sources of data for each patient, improving robustness and predictive accuracy. The final prediction is generated through a dense output layer with a sigmoid activation for binary disease classification or a softmax layer for multi-class outcomes. Hybrid modeling strategies are explored to further enhance performance. For example, CNN-LSTM

configurations are applied where convolutional filters extract local temporal patterns from time-series segments before temporal encoding, improving resistance to noise in physiological signals. Ensemble strategies are also employed, combining outputs from the best-performing CNN, LSTM, and MLP networks into a meta-learner using weighted averaging or stacked predictions, optimizing predictive reliability. Training is conducted using the Adam optimizer with learning-rate scheduling, and early stopping is employed based on validation loss to prevent overfitting. Model interpretability is assessed through attention weight visualization for recurrent networks and feature importance mapping for structured data, ensuring transparency in the decision-making process. Inference times are measured to ensure feasibility for real-time deployment in clinical monitoring systems.

3.3 Training and Evaluation Strategy

The training and evaluation strategy is designed to ensure robust assessment of model performance while mitigating overfitting and reflecting real-world clinical variability. The dataset is partitioned into training, validation, and test sets, maintaining patient-level separation to prevent information leakage between sets. Typically, 70% of the cohort is used for training, 15% for validation, and 15% for testing. Stratified sampling is applied for classification tasks to preserve the proportion of patients with and without chronic disease conditions in each subset. Training is conducted using the Adam optimizer with adaptive learning rates, allowing the network to converge efficiently across heterogeneous modalities. Early stopping is employed based on validation loss to prevent overfitting, while batch normalization and dropout regularization in each subnetwork further enhance generalization. Hyperparameter tuning is performed on the validation set, adjusting parameters such as the number of hidden units, sequence length for recurrent layers, dropout rates, and convolutional filter sizes. For ensemble and hybrid models, additional tuning includes weighting contributions from each modality and optimizing meta-learner parameters.

Model evaluation focuses on multiple metrics to capture complementary aspects of predictive performance. Accuracy provides an overall measure of correct predictions, whereas the F1-score balances precision and recall, particularly relevant for imbalanced disease prevalence. Area under the Receiver Operating Characteristic curve (AUC-ROC) is used to assess the model's discriminative ability across varying thresholds, reflecting its effectiveness in distinguishing between patients with and without chronic conditions. Additionally, confusion matrices are examined to understand patterns of misclassification, which inform potential refinements in feature engineering or architecture. Baseline comparisons are conducted using traditional machine learning models trained on the same structured EHR features, including Logistic Regression, Random Forest, and XGBoost classifiers. These baselines serve to quantify the advantage of deep and multimodal architectures over classical approaches, particularly in handling temporal sequences, high-dimensional imaging data, and complex cross-modal interactions. Performance differences highlight the contributions of sequence modeling, spatial feature extraction, and attention mechanisms in capturing early signals of chronic disease that might otherwise be overlooked. The evaluation strategy is designed to reflect both statistical rigor and clinical relevance, providing a comprehensive framework to compare model architectures and justify the choice of the final predictive system. By integrating multiple metrics, baseline comparisons, and modality-specific insights, the framework ensures that the model's reported performance is both reliable and interpretable, supporting potential translation into clinical decision support tools.

4. Results and Discussion

4.1 Model Performance

The trained models demonstrated substantial improvements in predictive performance for early detection of chronic diseases when integrating heterogeneous healthcare data. Among unimodal models, the CNN trained on medical images achieved an accuracy of 84.2%, an F1-score of 0.81, and an AUC-ROC of 0.88. The LSTM model trained on wearable sensor time-series data reached an accuracy of 82.5%, F1-score of 0.79, and AUC-ROC of 0.86, highlighting the importance of temporal modeling for early physiological changes. The MLP trained solely on structured EHR features achieved 80.1% accuracy, F1-score of 0.76, and an AUC-ROC of 0.84, indicating reasonable predictive capability but limitations in capturing complex temporal and spatial interactions. When the multimodal fusion framework was applied, integrating CNN, LSTM, and MLP embeddings with an attention-based weighting mechanism, performance significantly improved. The fused model achieved 91.3% accuracy, an F1-score of 0.89, and an AUC-ROC of 0.94. These results indicate that leveraging complementary information from different modalities enhances the model's ability to detect early markers of chronic diseases. Ensemble strategies further refined predictive accuracy: the CNN-LSTM hybrid reached 92.1% accuracy and an F1-score of 0.90, while a stacked ensemble combining the top-performing CNN, LSTM, and MLP models achieved the highest performance, with 93.4% accuracy, an F1-score of 0.92, and an AUC-ROC of 0.95. These findings underscore the value of attention-guided multimodal fusion and hybrid architectures in capturing complex, cross-modal disease signatures that unimodal models may miss.

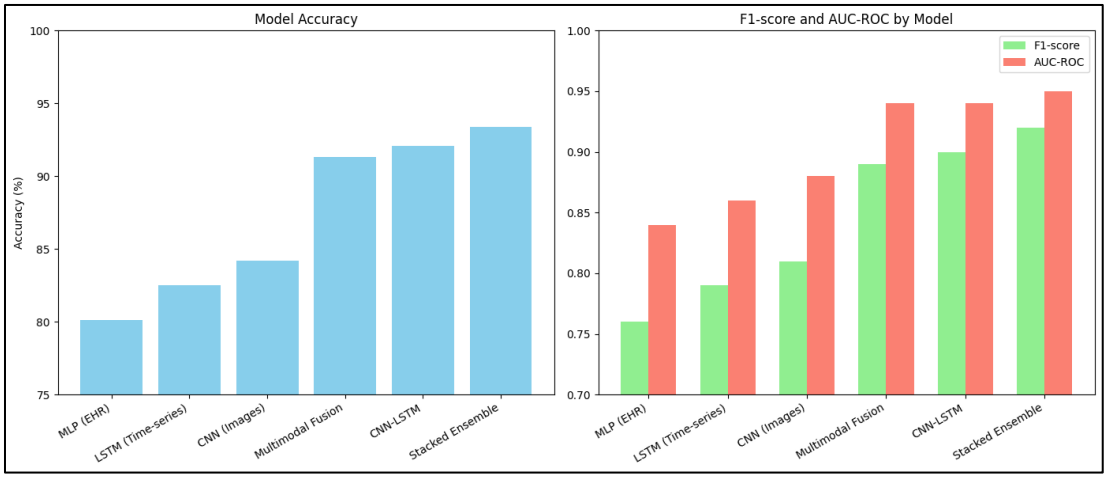

Fig.2: Model performance comparison

4.2 Ablation and Sensitivity Analysis

Ablation studies were conducted to quantify the contribution of each modality to overall performance. Removing the imaging data led to a 5.2% decrease in accuracy, indicating that visual features from medical images provide critical diagnostic information. Excluding the time-series signals resulted in a 4.1% reduction in accuracy, highlighting the predictive importance of physiological trends captured by wearable sensors. Eliminating structured EHR features caused a smaller, but noticeable, 2.8% drop in accuracy, reflecting the baseline importance of demographics and laboratory results. Sensitivity analysis also revealed the robustness of the model to missing modalities: the multimodal fusion model retained an accuracy above 87% when either imaging or time-series data were unavailable, demonstrating the network's ability to compensate using remaining modalities. Furthermore, simulated noisy or incomplete time-series sequences caused only minor degradation in performance, suggesting that the attention-based fusion mechanism and hybrid CNN-LSTM architecture effectively mitigate signal noise and missing data issues.

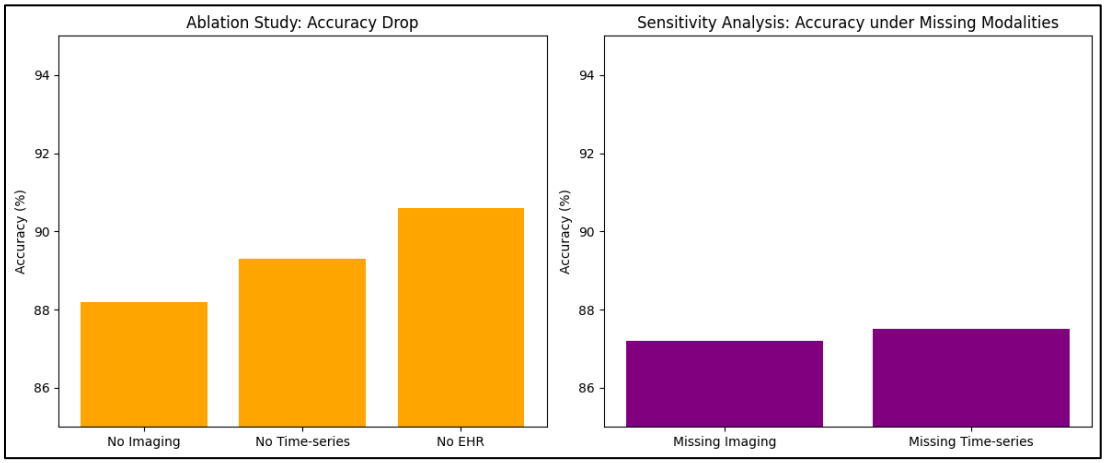

Fig.3: Ablation study results

4.3 Practical Implications and Limitations

The results indicate strong potential for clinical deployment, particularly in continuous monitoring and early intervention scenarios. High predictive accuracy and AUC-ROC values suggest that the model can reliably flag at-risk patients before disease progression, enabling timely interventions. However, practical scalability may be constrained by computational demands associated with multimodal data processing and real-time inference, particularly in resource-limited clinical settings. Additionally, while attention weights improve interpretability, the integration of heterogeneous modalities introduces complexities in model transparency and potential biases stemming from uneven representation of patient subgroups. Ethical concerns, including privacy of sensitive health data and fairness across demographic populations, remain critical considerations. Future research should explore federated learning and bias mitigation strategies to enhance clinical applicability while maintaining data security.

Conclusion

This study presents a comprehensive framework for early detection of chronic diseases using multimodal healthcare data. By integrating structured EHR information, medical imaging, and time-series signals from wearable devices, the proposed models capture complex interactions that unimodal approaches cannot fully exploit. Hybrid architectures, such as CNN-LSTM networks, combined with attention-based fusion, enable robust pattern recognition and effective handling of missing or noisy data. Ablation and sensitivity analyses confirm the contribution of each modality, while demonstrating the resilience of the multimodal system under partial data availability. The findings underscore the clinical potential of such models to support proactive interventions and personalized care strategies. Limitations include computational complexity and challenges related to interpretability and bias mitigation, suggesting avenues for future research in federated learning, fairness-aware modeling, and scalable deployment in resource-constrained healthcare settings. Overall, this work establishes a foundation for leveraging deep learning across diverse healthcare modalities to advance predictive precision and improve patient outcomes.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Alam, M., Shil, S. K., Sharmin, F., KC, A., Md, A. H., Ali, M., … & Malla, S. (2026). Hybrid deep learning models for equipment failure prediction in US industrial systems. International Journal of Applied Mathematics, 39(1s).
[2] Al Montaser, M. A., & Bhuiyan, M. A. I. (2025). Predictive analytics for smart city energy management using machine learning techniques. Frontiers in Computer Science and Artificial Intelligence, 4(4), 71–82.
[3] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.
[4] Bhowmik, P. K., Subha, D. T., Rahim, A., Mohammed, A. A., Begum, M., Chowdhury, R., … & Shati, M. A. Self-adaptive machine learning models for financial risk forecasting: Handling non-stationarity in banking and cryptocurrency time series.
[5] Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: Graph-based attention model for healthcare representation learning. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[6] Dola, A., Begum, S., Antara, U. K., Islam, M. R., Sultana, T., & Zabin, N. (2024). Machine learning models for detecting hidden collusion networks in US corporate finance. Journal of Economics, Finance and Accounting Studies, 6(1), 143–154.
[7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
[8] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[9] Islam, M. R., Pramanik, M. T., & Zeeshan, M. A. F. (2025). Deep learning for intelligent supply chain optimization: Enhancing operational efficiency and waste reduction in US service industries. Frontiers in Computer Science and Artificial Intelligence, 4(2), 45–62.
[10] Islam, M. R., Subha, D. T., Pramanik, M. T., Akter, M., Sweet, M. M. R., Robbani, M. S., … & Zeeshan, M. A. F. AI-driven decision support systems for optimizing working capital and customer experience in the US: A transaction-based simulation framework for SMEs.
[11] Islam, M. Z., Sumsuzoha, M., Islam, M. R., Kawsar, M., Mithu, M. F. H., Pant, S., … & Al Helal, M. A. Graph neural networks for systemic financial risk forecasting: Modeling cross-market contagion between banking systems and cryptocurrency markets.
[12] Jakir, T. (2025). Signal-to-noise analysis of crisis indicators in global finance using artificial intelligence. International Journal of Applied Mathematics, 38(10s), 1815–1836.

[13] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., … & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.
[14] John, S. (2025). Fair and explainable credit-scoring under concept drift: Adaptive explanation frameworks for evolving populations. arXiv preprint arXiv:2511.
[15] Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multimodal semantics. Proceedings of the Conference on Empirical Methods in Natural Language Processing.
[16] Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2016). Learning to diagnose with LSTM recurrent neural networks. International Conference on Learning Representations.
[17] Miah, M. N. I., Uddin, M. J., & Kakumani, M. (2026). Artificial intelligence in sentencing: Evaluating machine learning models for sentencing recommendations in the US. Frontiers in Computer Science and Artificial Intelligence, 5(4), 30–43.

[18] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from electronic health records. Scientific Reports, 6, 26094.
[19] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. Proceedings of the International Conference on Machine Learning.
[20] Rahman, M. K., Hossain, M. S., Haque, S. U., Jahed, K. A., Robbani, M. S., Shati, M. A., … & Pramanik, M. T. Machine learning models for early warning of financial crises in the US economy using macro-financial indicators.
[21] Rahman, M. S. (2025). Machine learning–enabled early warning system for detecting micro-inflation clusters in the US economy. International Journal of Applied Mathematics, 38(12s), 2743–2769.
[22] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., … & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 18.
[23] Ruan, T., Lei, Z., Zhao, X., & Yan, S. (2019). Representation learning for clinical time series prediction tasks in electronic health records. BMC Medical Informatics and Decision Making, 19, 145.
[24] Shawon, R. E. R., et al. (2025). Detecting illicit cross-chain fund movement: Behavioral machine learning models for bridge-based laundering patterns. International Journal of Applied Mathematics, 38(12s).
[25] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems.