RADemics

Data Acquisition and Preprocessing Techniques in AI Based Biomedical Systems

Satish Kumar Das, A. Perumal, A. Thanikasalam

RAJIV GANDHI UNIVERSITY, DOIMUKH, SETHU INSTITUTE OF TECHNOLOGY, ACADEMY OF MARITIME EDUCATION AND TRAINING DEEMED TO BE UNIVERSITY

Data Acquisition and Preprocessing Techniques in AI Based Biomedical Systems

[1]Satish Kumar Das, Assistant Professor, Department of Computer Science and Engineering, Rajiv Gandhi University, Doimukh, Arunachal Pradesh – 791112. satish.das@rgu.ac.in

[2]A. Perumal, Professor, Department of Mechanical Engineering, Sethu Institute of Technology, Pulloor, Karaiapatti, Virudhunagar-626115. perumalmech08@gmail.com

[3]A. Thanikasalam, Assistant Professor, Department of Marine Engineering, Academy of Maritime Education and Training, Deemed to be University, Kanathur, Chennai – 603112. thanikasalama@ametuniv.ac.in

Abstract

The increasing integration of artificial intelligence (AI) in biomedical systems has revolutionized healthcare, enabling more accurate diagnostics, personalized treatment, and real-time patient monitoring. However, the success of these AI models heavily depends on the quality and reliability of the underlying data. Biomedical data, which is often collected from a variety of sources such as medical imaging, wearable devices, and electronic health records (EHRs), is prone to noise, errors, and missing values. These issues can significantly impact the performance of AI models, leading to biased outcomes and unreliable predictions. This chapter delves into the critical aspects of data acquisition, preprocessing, and quality assurance techniques essential for enhancing the accuracy of AI-based biomedical systems. Key topics covered include advanced imputation methods for handling missing data, noise reduction techniques in medical imaging, dimensionality reduction strategies for high-dimensional datasets, and real-time data validation approaches for wearable devices. Special attention is given to the application of statistical models like Markov chains and Bayesian networks for error detection in sensor data. Emphasis is placed on the importance of effective data cleaning and preprocessing to mitigate overfitting, improve model robustness, and ensure the integration of high-quality data in AI workflows. The chapter concludes with a discussion on the future directions for developing more advanced, adaptive, and scalable data processing systems in the context of AI-driven healthcare solutions.

Keywords: Artificial Intelligence, Biomedical Data, Data Preprocessing, Real-Time Validation, Noise Reduction, Missing Data Imputation.

Introduction

Artificial intelligence (AI) has revolutionized healthcare by enabling faster, more accurate diagnoses, personalized treatments, and continuous patient monitoring [1]. From medical imaging to genomic analysis and wearable devices, AI systems are increasingly used to assist healthcare professionals in making data-driven decisions [2]. However, the success of these AI models heavily depends on the quality of the data they are built upon [3]. Biomedical data, which includes diverse types such as clinical records, medical imaging, sensor data, and genomics, often comes with a high level of complexity, noise, and incompleteness [4]. This presents significant challenges in ensuring that AI systems perform reliably and produce meaningful results. Effective data preprocessing and quality assurance are crucial in addressing these challenges, ensuring that AI models are fed with accurate, relevant, and high-quality data [5].

Data preprocessing plays a pivotal role in improving the performance of AI models [6]. Raw biomedical data typically suffers from issues such as missing values, noise, and inconsistencies [7]. For instance, medical images may be corrupted by artifacts, sensor data may have missing readings, and patient records may contain incomplete or erroneous entries [8]. In these cases, the data must undergo rigorous cleaning and validation processes to correct errors and fill in missing values. Techniques such as imputation methods, noise reduction algorithms, and dimensionality reduction strategies help transform raw data into a form suitable for training robust AI models [9]. Without such preprocessing steps, AI algorithms risk making predictions based on unreliable or incomplete data, leading to erroneous outcomes that could have severe consequences in healthcare [10].

One of the most critical challenges in AI-based biomedical systems is dealing with missing data [11]. Missing data is a common occurrence in healthcare, whether due to patient non-response, sensor malfunctions, or incomplete records [12]. The impact of missing data can be significant, leading to biased or incorrect model predictions if not properly handled [13]. Traditional methods, such as simply discarding incomplete records or imputing missing values with statistical averages, can result in information loss or inaccurate data imputation [14]. More advanced techniques, such as multiple imputation, k-nearest neighbor imputation, and deep learning-based autoencoders, offer more sophisticated approaches for dealing with missing data. These methods consider the underlying relationships within the data to produce more accurate imputations, ensuring that the missing values do not compromise the integrity of the AI model [15].