Muhammad E. H. Chowdhury
Serkan Kiranyaz *Editors*

Surveillance, Prevention, and Control of Infectious Diseases

An AI Perspective

Springer

Surveillance, Prevention, and Control of Infectious Diseases

Muhammad E. H. Chowdhury • Serkan Kiranyaz
Editors

Surveillance, Prevention, and Control of Infectious Diseases

An AI Perspective

Springer

Editors

Muhammad E. H. Chowdhury
Department of Electrical Engineering
College of Engineering, Qatar University
Doha, Qatar

Serkan Kiranyaz
Department of Electrical Engineering
College of Engineering, Qatar University
Doha, Qatar

ISBN 978-3-031-59966-8 ISBN 978-3-031-59967-5 (eBook)
https://doi.org/10.1007/978-3-031-59967-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Infectious diseases pose a significant threat to public health worldwide, requiring constant vigilance, innovative strategies, and rapid responses to mitigate their impact. In recent years, Artificial Intelligence (AI) has emerged as a powerful tool in the fight against infectious diseases, revolutionizing the way we surveil, prevent, and control outbreaks. This book, *Surveillance, Prevention, and Control of Infectious Diseases: An AI Perspective*, explores the intersection of AI and infectious disease management, offering insights into the latest advancements, challenges, and opportunities in this rapidly evolving field.

The chapters in this book provide a comprehensive overview of various aspects of infectious disease management through the lens of AI. We begin with an exploration of the prospects and challenges of using AI for infectious disease detection, highlighting its potential to expedite outbreak identification and facilitate proactive public health interventions. We then delve into the development and application of AI algorithms for rapid pathogen detection and surveillance, discussing innovative approaches for identifying and monitoring infectious diseases in real-time. One of the key areas covered in this book is the use of wearable devices and continuous physiological signal monitoring for early disease detection. We examine how wearable sensors and AI algorithms can be utilized to analyze physiological data and identify early symptoms of infectious diseases, enabling timely intervention and improved patient outcomes. Additionally, we explore the development of low-cost, automated digital microscopes for detecting diseases such as malaria, showcasing the potential of AI-driven technologies to improve diagnostic capabilities in resource-limited settings. This book addresses the challenge of under-vaccination in developing countries, proposing a comprehensive framework to identify and support children who miss essential vaccinations. It reviews research on vaccine gaps, synthesizes effective intervention strategies, and explores the potential of advanced technologies like AI and blockchain. By offering a systematic approach, the framework aims to improve vaccination coverage and promote global health equity by ensuring equitable access to vaccines for all children. Another highlight of this book is the discussion on the interpretability of deep learning models for tuberculosis detection using X-ray images. We explore how lightweight parallel

CNN models can enhance diagnostic accuracy while providing insights into the prediction process, paving the way for more effective disease diagnosis and management. Furthermore, we investigate the use of AI in the surveillance of seasonal respiratory infections, showcasing its role in automating outbreak identification and reducing transmission rates through early detection and intervention.

Throughout this book, we emphasize the importance of collaboration between healthcare professionals, researchers, policymakers, and technologists in harnessing the full potential of AI for infectious disease management. We hope that the insights and findings presented in this book will inspire further research and innovation in this critical area, ultimately leading to more effective strategies for surveilling, preventing, and controlling infectious diseases on a global scale.

We are immensely grateful for the dedicated efforts of our contributors in shaping this comprehensive volume on *Surveillance, Prevention, and Control of Infectious Diseases: An AI Perspective*. Your expertise and commitment have ensured the quality and relevance of each chapter, enriching the book with diverse insights and innovative approaches. As we move forward with the editing process, we encourage thoroughness and attention to detail to maintain the book's excellence. We extend our heartfelt appreciation to our esteemed reviewers for their invaluable feedback and constructive criticism, which have been instrumental in refining the manuscript. Your thorough reviews have enhanced the rigor and clarity of the content, contributing significantly to the scholarly integrity of the book. We are deeply grateful for your time, expertise, and dedication to advancing knowledge in this critical field.

Doha, Qatar

Muhammad E. H. Chowdhury
Serkan Kiranyaz

About the Book

Surveillance, Prevention, and Control of Infectious Diseases: An AI Perspective is a cutting-edge book that explores the transformative role of Artificial Intelligence (AI) in managing infectious diseases. Through ten comprehensive chapters, the book delves into various facets of AI application, offering insights into disease surveillance, prevention, and control. Beginning with an introduction to the potential of AI in infectious disease detection, the book progresses to review AI algorithms for rapid pathogen detection and surveillance. It further examines the integration of big data into infectious disease surveillance and discusses the development of low-cost, automated digital microscopes for malaria detection. The book also explores innovations in tuberculosis screening, robust dengue patient detection, interpretable deep learning models for tuberculosis detection, and continuous physiological signal monitoring using wearables for early infectious disease detection. Additionally, it discusses AI's role in monitoring seasonal respiratory infections, highlighting its contributions to detection, forecasting, and management. By showcasing the efficacy of AI-driven approaches in disease identification, the book underscores their potential to revolutionize infectious disease management, making it an indispensable resource for healthcare professionals, researchers, and policymakers striving for more effective disease surveillance and control strategies.

Contents

Artificial Intelligence for Infectious Disease Detection: Prospects and Challenges

Md. Moradul Siddique, Md. Masrafi Bin Seraj, Md. Nasim Adnan, and Syed Md. Galib

1 Introduction

Infectious diseases are sicknesses brought on by pathogens like bacteria, viruses, fungi, or parasites that can spread from person to person or by transporters like mosquitoes. It poses a significant threat to global health, causing widespread morbidity, mortality, and socioeconomic disruption. Infectious diseases can range from common infections like the flu to more severe outbreaks like COVID-19 and historical pandemics like the Black Death [91]. From an AI perspective, infectious diseases present a complex challenge that can benefit from advanced technological solutions. AI offers the potential to enhance various aspects of infectious disease management, including early detection, accurate diagnosis, outbreak prediction, and treatment optimization [85]. By leveraging machine learning algorithms, deep neural networks, and data analytics, AI can process vast amounts of medical data, identify patterns, and provide insights that empower healthcare professionals and

M. M. Siddique
Department of Computer Science and Engineering, Jashore University of Science and Technology (JUST), Jashore, Bangladesh

Artificial Intelligence for Development Lab (AID Lab), Jashore University of Science and Technology, Jashore, Bangladesh

Department of Computer Science and Engineering, University of Information Technology & Sciences (UITS), Dhaka, Bangladesh

M. M. B. Seraj · M. N. Adnan · S. M. Galib (✉)
Department of Computer Science and Engineering, Jashore University of Science and Technology (JUST), Jashore, Bangladesh

Artificial Intelligence for Development Lab (AID Lab), Jashore University of Science and Technology, Jashore, Bangladesh
e-mail: galib.cse@just.edu.bd

M. E. H. Chowdhury, S. Kiranyaz (eds.), *Surveillance, Prevention, and Control of Infectious Diseases*, https://doi.org/10.1007/978-3-031-59967-5_1

public health authorities to respond effectively to outbreaks and minimize their impact on global health.

In terms of timely response, early detection allows for a rapid response to contain and mitigate the spread of the disease. Swift identification of cases enables health authorities to implement measures such as isolation, quarantine, and treatment, reducing the chances of widespread transmission [83]. Hospitals and healthcare facilities can prepare for potential surges in patient numbers, ensuring that sufficient medical supplies, personnel, and equipment are available to provide adequate care. Swift detection facilitates the identification of the source of the outbreak, helping to trace and isolate cases before they spread widely. This is particularly crucial in infectious diseases with high transmission rates. In addition, early detection contributes to robust disease surveillance systems. Monitoring the progression of an outbreak provides insights into its dynamics, allowing for adjustments to control strategies as needed [86].

However, the journey towards a harmonious integration of AI in infectious disease detection is not without its share of challenges. Data quality, availability, and privacy emerge as formidable hurdles, demanding innovative strategies to ensure a robust and representative dataset. Ethical considerations of patient confidentiality, consent, and transparency underscore the need for responsible AI deployment in a domain where human lives are at stake [89].

Certainly, there have been several successful applications of AI in various medical domains that demonstrate its potential for revolutionizing healthcare. These successes set the stage for AI's promising role in infectious disease detection and management. Remarkably, Medical Imaging Analysis, Drug Discovery and Development, Personalized Medicine, Patient Care and Management, Natural Language Processing (NLP) in healthcare, and Robot-Assisted Surgery are some of the successful and ongoing AI-based techniques currently in use in disease detection and management [84].

Interdisciplinary collaboration between AI researchers, healthcare professionals, and policymakers is paramount in harnessing the full potential of AI in healthcare. This collaboration brings together diverse expertise and perspectives to address complex challenges, drive innovation, and ensure the responsible and effective integration of AI technologies. Apart from the chapter navigates through the diverse terrain of AI applications in infectious diseases detection, spotlighting successful cases. Whereas AI's prowess has led to early warnings of outbreaks, accurate diagnostics through medical imaging analysis, and proactive management of disease spread. As we navigate the dynamic interplay between technology and medicine, the chapters to come will elucidate the path towards a future where AI and human expertise synergize to fortify our defenses against infectious diseases in an ever-evolving global landscape.

2 Background Study

Our aim is to shed light on major Artificial Intelligence approaches for infectious disease detection. Therefore, we will discuss some preliminaries in this section.

2.1 Artificial Intelligence

Artificial Intelligence comprises various technologies, including machine learning, deep learning, natural language processing, and data analytics. These tools empower computers to execute activities that traditionally demand human intelligence, like identifying patterns, forecasting, and gaining knowledge through experience. In healthcare, AI is being harnessed to tackle some of the most pressing challenges, including improving diagnostics, personalizing treatment plans, managing medical records, and even drug discovery [46]. Early detection and accurate diagnosis play a pivotal role in effectively controlling infectious disease outbreaks. Whereas AI systems can continuously learn from new data, improving their detection and diagnostic capabilities over time. This adaptability ensures that the AI remains effective even as pathogens evolve, or new infectious diseases emerge.

2.2 Machine Learning

Machine Learning (ML) constitutes a subset of AI that centers on creating algorithms and models to facilitate computers in acquiring knowledge from data, thereby enabling them to formulate forecasts and choices. The fundamental idea is that machines can automatically improve their performance on a task through experience. A graphical illustration of the disease detection framework using machine learning classification techniques is shown in Fig. 1.

ML algorithms require substantial amounts of data to learn patterns and relationships. This data can be labeled (with known outcomes) or unlabeled, and it serves as the foundation for training models. Later on, features are the attributes or characteristics extracted from the data that the model uses to make predictions.

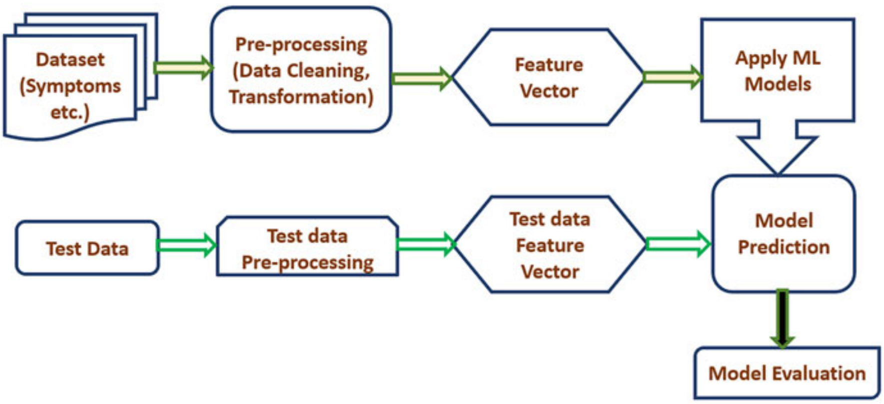

Fig. 1 A framework for disease detection method using Machine Learning

Effective feature selection and engineering are crucial for model performance. During training, the model learns patterns and relationships in the data by adjusting its parameters. The model is exposed to a dataset and learns to make predictions based on the input features [88]. The algorithms used in machine learning can be categorized into two major categories—supervised and unsupervised learning models. In a supervised learning model, the model learns from labeled data, where the input is paired with the correct output. It aims to learn the mapping between inputs and outputs to make predictions on new, unseen data [87]. On the other hand, unsupervised learning deals with unlabeled data. The goal is to uncover hidden patterns or structures within the data, such as clustering similar data points together [90].

2.3 Deep Learning

Deep Learning (DL) stands as a subdivision of machine learning, emphasizing the utilization of multi-layered neural networks to simulate and address intricate challenges. These neural networks draw inspiration from the human brain's architecture, composed of interconnected nodes (neurons) arranged in layers. Particularly, deep neural networks encompass numerous layers of interconnected nodes. Each layer extracts and transforms features from the input data [52]. Deep learning leverages hierarchical feature representations, with early layers detecting simple features and later layers combining them to detect more complex patterns. For instance, CNNs are particularly effective for image and visual data. They use convolutional layers to automatically learn relevant features from images [2]. Based on the image data, a conventional framework of a deep learning approach (CNN) is shown in Fig. 2. On the flip side, RNNs are suited for sequential data, like time series or natural language. They have connections that loop back, allowing them to capture temporal dependencies [50].

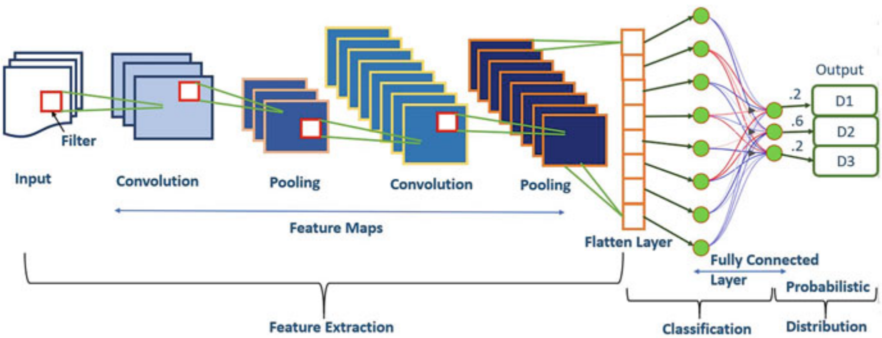

Fig. 2 A general framework of the Convolutional Neural Network (CNN)

In a deep neural network (e.g., a CNN in Fig. 2), the first few layers capture simple features like edges, textures, and basic shapes from the raw input data. As information propagates through subsequent layers, these simple features are combined to recognize more complex structures, such as object parts or specific characteristics. Each layer in the network performs a transformation on the input it receives, generating a new representation that highlights specific aspects of the data. This process is referred to as feature learning [43]. Deeper layers learn to represent high-level features that are combinations of the lower-level features learned by preceding layers. Furthermore, deep networks typically use non-linear activation functions (e.g., ReLU, sigmoid) between layers. These activation functions introduce non-linearity, enabling the network to capture intricate relationships in the data that linear models cannot easily handle [49]. As a result, deep networks allow end-to-end learning, where the entire system learns directly from raw input data to produce desired outputs. This eliminates the need for manual feature engineering and intermediate processing steps, making the modeling process more streamlined. One challenge of deep networks is the vanishing or exploding gradient problem, which can affect the stability of training [64]. Several techniques like skip connections and batch normalization have been introduced to mitigate these issues, enabling the training of very deep networks [31]. To sum up, DL focuses on learning feature representations directly from data, eliminating the need for manual feature engineering in many cases.

2.4 The Intersection of AI and Infectious Diseases

Artificial Intelligence (AI) can significantly complement traditional methods of infectious disease detection by providing enhanced speed, accuracy, and scalability [41]. AI algorithms can rapidly process large datasets, such as patient records, medical images, and genetic sequences. This speed enables quicker identification of patterns and anomalies associated with infectious diseases. As a consequence, AI algorithms are capable of identifying subtle patterns and correlations that may go unnoticed by human experts, leading to more accurate disease detection and diagnosis [48]. Besides, AI systems can handle an enormous volume of data simultaneously, making them well-suited for processing data from diverse sources and at different scales. In terms of genomic data, AI can swiftly analyze genomic data to identify specific genetic markers associated with infectious diseases, aiding in early detection and personalized treatment strategies [21]. Meanwhile, AI algorithms can uncover genetic variations and mutations that might contribute to disease transmission or resistance, guiding effective intervention measures [23]. AI accelerates drug discovery by analyzing molecular structures, identifying potential drug candidates, and predicting their interactions with pathogens [57]. By complementing traditional methods with AI-powered techniques, infectious disease detection can become more efficient, accurate, and scalable, ultimately leading to better preparedness and response in addressing health crises.

3 Related Work

In disease detection, several machine learning algorithms were used, each with strengths and effectiveness for particular types of data and disease identification tasks. In supervised learning, Support Vector Machine (SVM), Random Forests (RF), Naïve Bayes (NB), Decision Tree (DT), etc. can be used for early disease diagnosis by learning patterns from patient data and classifying new cases [8].

In order to forecast disease outbreaks, researchers utilize machine learning algorithms to assess epidemiological data, climate information, and population mobility. For instance, [67] used machine learning to forecast dengue fever outbreaks based on historical incidence and climate factors. AI-powered wearable devices and remote monitoring systems can track patient vital signs and symptoms, enabling healthcare providers to remotely monitor and manage infectious disease cases [7].

The pandemic of coronavirus disease 2019 (Covid-19) has demonstrated the ease of how quickly illnesses can spread and jeopardize the world [62, 63]. The primary focus on advancing neural network methods for COVID-19 detection resulted in the creation of numerous cutting-edge neural network models aimed at improving the accuracy of COVID-19 detection. Researchers developed deep learning models, including Convolutional Neural Networks (CNNs), Transfer Learning to analyze chest X-rays and CT scans for COVID-19 detection [37]. In essence, AI models assist radiologists in identifying COVID-19-related patterns in lung images, aiding in early diagnosis and monitoring disease progression [3, 59].

Parkinson's disease is a neurological condition that worsens over time and impairs movement control. Detecting Parkinson's disease through AI involves utilizing machine learning algorithms and data analysis techniques to analyze various data sources and identify patterns indicative of the disease. In Parkinson's disease (PD) detection, researchers used different types of data such as movement data, voice and speech data, imaging data (MRI or PET scans), and clinical data [82].

To detect red blood cells infected with malaria, machine learning algorithms were trained using microscopic blood smear images [54]. When compared to other diseases, heart disease (HD) affects the lives of the greatest number of people worldwide. With the advent of advanced machine learning techniques, researchers have been able to develop predictive models that aid in the early detection of heart disease, thereby potentially saving lives [16]. In order to predict the spread of the Zika virus, researchers developed machine learning models by analyzing factors like mosquito density, climate, and human mobility [35]. Due to the virus's highly rapid transmission and potential to infect people indiscriminately, researchers employed machine learning algorithms on the disease dataset to aid in predicting and preparing for the influenza pandemic [6, 38].

Diabetes has become a growing worry due to its elevated morbidity rates, and the age range of those impacted by this condition has now shifted to the early twenties. Numerous researchers and medical professionals have currently devised methods of detection utilizing artificial intelligence, aiming to address issues that may go unnoticed due to human mistakes [18, 70].

The recent occurrence of monkeypox (MPX) infection has evolved into a worldwide issue of significant interest over the previous year. Epidemiological studies employing artificial intelligence for managing the ongoing MPX outbreak need to account for the origin of the infection and assess all avenues of transmission [15].

Every year, dengue's notable increase constitutes 10% of fevers in youth within endemic nations. Given similar symptoms to other viruses, early diagnosis challenges persist, potentially exacerbated by inadequate sensitive tools. In upcoming diagnostic approaches, a combination of viral and clinical indicators will be essential, sequentially analyzed using artificial intelligence technology. This will enable a more comprehensive assessment of illness onset, severity determination, and treatment management [11, 77].

Pneumonia is a prevalent respiratory infection that poses a potential threat to lives, impacting millions of individuals worldwide. Recently, the incorporation of AI methods has displayed encouraging outcomes in improving the identification of pneumonia [75].

Even with the progress achieved in expanding HIV treatment initiatives, there still exists a significant unfulfilled need for monitoring disease progression and treatment efficacy, posing a challenge to HIV/AIDS management. It can be conducted evaluations of viral load and CD4 classification in adults receiving ART care, employing machine learning algorithms [47]. Table 1 represents a comprehensive study focusing on infectious diseases through the lens of artificial intelligence.

4 Methodology

The section encompasses a diverse set of components, encompassing various types of data, techniques for data pre-processing, methods for feature extraction, and the specific focus on the detection or classification of infectious diseases.

(a) *Data and Data Types*

AI algorithms can analyze vast amounts of data from various sources, such as social media, healthcare records, and environmental data. This reduces the chances of misdiagnosis and ensures that infected individuals receive appropriate treatment promptly. By continuously monitoring this data, AI can detect patterns and anomalies that may indicate the emergence of a new infectious disease or the spread of an existing one much faster than traditional methods [69].

For medical images data such as X-rays, computed tomography (CT) scans, and magnetic resonance imaging (MRI) provide visual insights into disease progression, localization of infections, and organ damage [17]. Using social media and internet data, digital surveillance can provide early signals of disease

Table 1 A comprehensive study of the infectious disease based on AI

Reference	Problem area	Data type	Data size	Techniques	Results
Ahishali et al. [3]	COVID-19 diagnosis	X-ray images	1065 samples (pneumonia with COVID-19 and 12,544 samples (normal)	DenseNet-121	Accuracy of 99.2%, sensitivity 95% and specificity 99.74%
Yamac et al. [81]	COVID-19 recognition	X-ray images	2760 samples (bacterial pneumonia), 1485 samples (viral pneumonia), 462 samples (COVID-19) and 1579 normal samples	Convolution support estimation network (CSEN2)	Accuracy of 95.9%, sensitivity 98.5% and specificity 95.7%
Qiblawey et al. [59]	COVID-19 detection and severity	CT images	1139 patients with 51,027 CT samples	DenseNet201 FPN	Accuracy of 95.96%, precession of 96.47% sensitivity of 99.64%, specificity of 98.72%, and F1-score of 98.03%
Rahman et al. [61]	COVID-19 detection	Chest X-ray images	18,479 samples including 3616 COVID-19 samples	Pre-trained model (ResNet18, ResNet50, ResNet101, InceptionV3, DenseNet201, and ChexNet)	The highest accuracy is 96.29%, F1-score is 96.28% and sensitivity is 97.28%
Degerli et al. [19]	COVID-19 detection and severity grading	Chest X-ray images	119,316 where 2951 samples of COVID-19	U-Net DenseNet-121	Accuracy is 99.73%, precision 96.40%, F1-score 95.67%, sensitivity 94.96%, and specificity 99.88%.
Degerli et al. [20]	COVID-19 detection	Chest X-ray images	121,378 samples including 9258 COVID-19 samples	OSegNet (DenseNet-121 and inception-v3)	The maximum accuracy of 99:65% with 98:09% precision
Brioschi et al. [14]	COVID-19 detection	Thermal images	2558 cases from 227,261 workers in five countries	CNN	Accuracy of 98.98% and a PPV of 18.18%
Azeem et al. [10]	COVID-19 detection	Chest radiography images	10,192 benign 3616 malignant	ConXNet	Accuracy of 97.8%, precision is 97.93% and F1-measure is 97.92%

Lee et al. [44]	Malaria prediction	Blood smear images	Over 10,000 cases from 1956 to 2019	SVM, RF, MLP, AdaBoost, GB, and CatBoost	Accuracy of SVM: 93.8%, RF: 93%, MLP:91.4%, AdaBoost: 93.2%, GB: 93%, and CatBoost: 94.9%
Kumar et al. [42]	Malaria detection	Blood smear images	27,558 cases from 200 patients	CNN	Accuracy of 95.4%
Gan et al. [25]	Forecast Hepatitis B	Text data	10,486,959 cases	Hybrid model (BP-ANN and GM)	MAE: 3.9704 × 104, RMSE: 4.863 × 103, R: 0.9495
Harabor et al. [28]	Prediction of Hepatitis B and C	Clinical data	1359 patients	SVM, RF, NB, and KNN	The KNN achieves highest accuracy of 98.1%, SVM and RF with equal accuracies (97.6%) and NB (95.7%)
Doyle et al. [22]	Undiagnosed patient detection of Hepatitis C	Clinical data	9,721,923 patients	LR, GB trees, RF, and stacked ensemble	The highest AUCROC of 96%, specificity of 99.9%, and precision of 97%
Saleh and Rabie [66]	Monkeypox diagnose	Image data	MPX 296 samples and others 204 samples	NB, kNN, and DL	Accuracy: 98.48%, precision: 91.1%, and recall:88.91%
Islam and Shin [33]	Monkeypox detection	Image data	MPX 1428 samples and others 1764 samples	ResNet18 and FL	Accuracy of 99.81%
Turbé et al. [78]	Rapid HIV diagnostic	Image data	11,374 images	MobileNetV2, MobileNetV3, ResNet50, SVM	Sensitivity: 97.8%, specificity: 100%
Seboka et al. [68]	HIV/AIDS treatment activities	Demographics and clinical data	2907 adults	kNN, DT, NB, GNB, SVM, LR, GB, XGB, RF	Accuracy of 96%, sensitivity 97%, AUC: 0.99, F1-measure: 96%

outbreaks, public sentiment, and information dissemination [1]. Similarly, the genomic data of pathogens, obtained through techniques like DNA sequencing, help identify the specific microorganism responsible for an infectious disease [26]. Consequently, wearable devices and sensors can collect real-time data on physiological parameters like body temperature, heart rate, and respiratory rate [72]. Thus, AI algorithms can detect anomalies that might indicate the presence of an infectious disease.

In infectious disease detection, a diverse range of data types such as clinical data, imaging data, genomic data, social media and web data, sensor data, drug and treatment data, healthcare system data, etc. are harnessed to gain insights into the presence, spread, and characteristics of diseases. Each type of data offers unique information that enhances our ability to detect and manage infectious diseases effectively when combined and analyzed using advanced AI techniques.

(b) *Data preprocessing*

Data preprocessing involves preparing and cleaning raw data to ensure that it is suitable for analysis by AI algorithms. Also, data preprocessing aims to enhance the quality, reliability, and effectiveness of AI models. There are some key aspects of data preprocessing such as data cleaning (removing or correcting errors, inconsistencies, and missing values), data integration (combining data from different sources as well as ensuring data compatibility and standardization), normalization (bringing different features to a similar scale), data augmentation (increase the diversity of the dataset), handling imbalanced data, dimensionality reduction and so on [32, 56]. The traditional data preparation techniques are illustrated in Fig. 3.

(c) *Feature Extraction and Detection*

The feature extraction techniques involve transforming raw data into a set of relevant and informative features that capture the essential characteristics of the data. Feature extraction helps highlight patterns, anomalies, and relationships that can aid AI models in making accurate predictions. Usually, there are different types of features such as statistical features, text and Natural Language Processing (NLP) features, image and signal processing features, molecular and genetic features, composite features, and so forth [12]. Several methods of feature extraction are utilized in the identification of infectious diseases. In textual data, NLP techniques can be used to extract features like word frequencies, TF-IDF scores, n-grams, and topic modeling [24]. For image data, features might include textures, edges, shapes, color histograms, and moments that describe different aspects of the image [30]. In addition, pre-trained deep learning models, such as convolutional neural networks (CNNs) or transformer-based models, can be used to extract high-level features from data like images, text, or audio [40]. In addition, Cross-validation techniques may also be used to evaluate and tune the model's hyperparameters.

In traditional machine learning, feature extraction involves manually selecting and transforming specific aspects of the data to create meaningful inputs

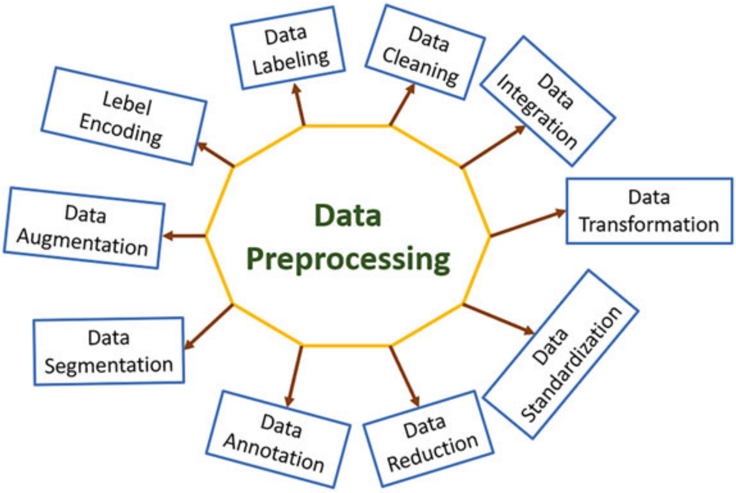

Fig. 3 Conventional data preprocessing techniques in AI

for the model. However, in deep learning, neural networks are designed to automatically learn hierarchical representations of data through layers of interconnected nodes (neurons). Each layer in a neural network learns to recognize increasingly complex patterns or features from the data. This ability to learn features internally is one of the key advantages of deep learning. For example, in an image classification task, traditional machine learning might require manually extracting features like edges, textures, or shapes from the images before feeding them into the model. In contrast, a deep learning model like Fig. 2, a convolutional neural network (CNN) can automatically learn to detect edges, textures, and other relevant features directly from the raw pixel values of the images. Similarly, in natural language processing tasks, deep learning models like recurrent neural networks (RNNs) or transformers can learn to understand the underlying semantic structure of text without requiring explicit manual feature engineering. This internal feature extraction capability of deep learning methods not only simplifies the model-building process but also often leads to improved performance, as the models can adapt and learn relevant features from the data on their own [43].

5 Performance Metrics

The Confusion Matrix is the common performance evaluation tool used in the field of machine learning and statistics to assess the performance of a classification model. The Confusion Matrix is a square matrix that summarizes the predictions

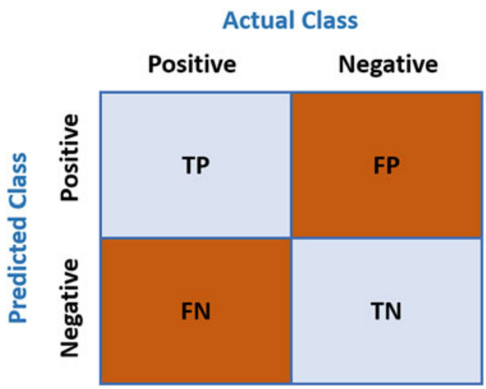

Fig. 4 Binary Confusion Matrix

made by a classification model on a set of data points with known ground truth labels. It provides a breakdown of correct and incorrect predictions made by the model for each class in the dataset. The matrix is of size N × N, where N is the number of classes in the dataset. For a binary classification problem, N will be 2, and for a multi-class problem, N will be greater than 2.

When N = 2, the general structure of a Confusion Matrix is shown in Fig. 4.

- True Positive (TP): The number of positive data points that were successfully categorized as positive (correctly forecasted as positive) outcomes.
- False Positive (FP): The number of negative data points that were incorrectly categorized as positive (or incorrectly forecasted as positive)
- True Negative (TN): The number of negative data points that were successfully categorized as negative (correctly forecasted as negative) outcomes.
- False Negative (FN): The number of positive data points that were incorrectly categorized as negative (incorrectly forecasted as negative) outcomes.

The aforementioned information is used to formulate many performance indicators that describe numerous aspects of how well the model performs [45]. From the Confusion Matrix, the following typical metrics can be derived:

Accuracy is an important performance parameter for assessing an AI model's effectiveness. It estimates the percentage of correctly forecast instances (or data points) among all of the instances (instances belonging to the class) in the dataset.

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

Accuracy can be misleading, especially when dealing with unbalanced datasets. An unbalanced dataset is one where the distribution of classes is heavily skewed, meaning that one class has significantly more examples than the other(s). In such cases, accuracy alone may not provide a clear and accurate picture of a model's performance. For example, if we have a dataset with 95% of instances belonging to Class A and only 5% to Class B, a model that predicts Class A for all instances would still achieve an accuracy of 95%, even though it is not actually making any meaningful predictions.

Recall, also known as Sensitivity or True Positive Rate, is a performance metric that focuses on the ability of a classification model to correctly identify all relevant instances of a particular class. When considered in terms of the confusion matrix, recall is calculated by considering the true positives (TP) and false negatives (FN) for a specific class.

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

Recall is particularly important when the consequences of missing positive instances (false negatives) are more severe than misclassifying negative instances (false positives). In other words, recall is useful when we want to minimize the risk of failing to identify instances of a certain class [80]. A high recall indicates that the model is effectively capturing most of the positive instances in the dataset, while a low recall suggests that the model is missing a significant number of positive instances. However, achieving high recall might come at the cost of increased false positives, which in turn reduces the precision.

Specificity or the True Negative Rate is a performance metric used in the evaluation of a classification model's ability to correctly identify the negative instances. It complements the concept of recall, which focuses on the model's ability to correctly identify positive instances. In other words, specificity is the recall rate of the negative class. Regarding the confusion matrix, specificity is calculated by considering the true negatives (TN) and false positives (FP) for the negative class. A high specificity indicates that the model is effective at correctly classifying negative instances, while a low specificity suggests that the model includes a considerable number of false positives in its negative predictions.

$$Specificity = \frac{TN}{TN + FP}$$

Precision Focuses on the accuracy of positive predictions made by a classification model. It provides a measure of how well the model is at avoiding false positive errors (false alarms). In terms of the confusion matrix, precision is calculated by considering the true positives (TP) and false positives (FP) for a specific class. A high precision indicates that the model is making accurate positive predictions, while a low precision suggests that the model is including a substantial number of false positives in its positive predictions [36].

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