Bulletin of Electrical Engineering and Informatics
Vol. 11, No. 6, December 2022, pp. 3509~3520
ISSN: 2302-9285, DOI: 10.11591/eei.v11i6.4225

Machine learning approaches in the diagnosis of infectious diseases: a review

Smriti Mishra[1], Ranjan Kumar[2], Sanjay Kumar Tiwari[3], Priya Ranjan[4]
[1]Department of Computer Science, Gaya College, Gaya, India
[2]Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India
[3]Department of Mathematics, Magadh University, Bodhgaya, India
[4]Department of Electrical and Electronics, Bhubaneswar Institute of Technology, Bhubaneswar, India

Article Info

Article history:

Received Jun 9, 2022
Revised Jul 25, 2022
Accepted Aug 22, 2022

Keywords:

Early diagnosis
Ensemble learning
Infectious diseases
Machine learning

ABSTRACT

Infectious diseases are a group of medical conditions caused by infectious agents such as parasites, bacteria, viruses, or fungus. Patients who are undiagnosed may unwittingly spread the disease to others. Because of the transmission of these agents, epidemics, if not pandemics, are possible. Early detection can help to prevent the spread of an outbreak or put an end to it. Infectious disease prevention, early identification, and management can be aided by machine learning (ML) methods. The implementation of ML algorithms such as logistic regression, support vector machine, Naive Bayes, decision tree, random forest, K-nearest neighbor, artificial neural network, convolutional neural network, and ensemble techniques to automate the process of infectious disease diagnosis is investigated in this study. We examined a number of ML models for tuberculosis (TB), influenza, human immunodeficiency virus (HIV), dengue fever, COVID-19, cystitis, and nonspecific urethritis. Existing models have constraints in data handling concerns such data types, amount, quality, temporality, and availability. Based on the research, ensemble approaches, rather than a typical ML classifier, can be used to improve the overall performance of diagnosis. We highlight the need of having enough diverse data in the database to create a model or representation that closely mimics reality.

Corresponding Author:

Ranjan Kumar
Department of Computer Science, Aryabhatta College, University of Delhi
Benito Juarez Road, New Delhi, Delhi-110021, India
Email: ranjan301@gmail.com

1. INTRODUCTION

Infection is characterized as the invasion of an organism within human body tissues by disease-causing agents. It starts growing within the host and releases toxins. Infections are caused by infectious agents known as pathogens, which include bacteria, viruses, fungi, protozoa, and parasites. These parasites can consist of unicellular organisms like the malaria parasite and macro parasites like the worm's cyst [1]. A disease can be infectious if the causing agent is a pathogen. An infectious disease, also known as transmissible disease or communicable disease, is an illness resulting from an infection. Pathogen reproduction in the host causes a variety of interferences, ranging from membrane rupture in viruses to the release of toxins in many bacterial diseases. These interfering signals activate molecular biological techniques to protect cells from intruders and warn them of potential threats. These infections can also modify pathways and cause long-term organ damage, such as lung tissue scarring resulting from immune hyperactivity. Thus, the same pathways intended to reduce and prevent this infection can have relatively long implications. These pathogens can be transmitted to others either directly or indirectly. They can be influenced by various environmental factors such as infrastructure, land use changes, travel, and commerce, natural disasters, climate, war and conflict, and evolving technologies and industries. Table 1 summarizes the transmission of infectious diseases.

Table 1. Infectious disease spread: direct person–to–person, indirect person-to-person, common vehicle spread, zoonosis, and vector-borne

Direct person-to-person	Indirect person-to-person	Common vehicle spread	Zoonosis	Vector-borne
– Sexual transmission – Needle injection – Skin-to-skin – Human bites – Perinatal mother to child transmission	– Contaminated objects	– Food borne, – Waterborne, – Fecal-oral	– Animal bytes – Air borne – Blood borne	– Mosquitoes, – Flies, – Fleas, – Ticks

Infectious diseases are worldwide issues that are a major cause of death. Infectious disease detection and diagnosis are always a key concern in public health or the economy. Traditionally, this has been accomplished by tracing the factors, roots, or pathways of infectious agent transmission, and identifying patterns [2]. If an early diagnosis is not possible, serious complications or death may result. Epidemic diseases are an unexpected growth in the percentage of disease cases in a limited geographical area. It usually grows linearly. The World Health Organization declares a disease pandemic when the growth rate is exponential, and the disease spreads across countries.

Artificial intelligence (AI) is a technique to create computer systems capable of learning "patterns”, AI has grown exponentially in the last decades, resulting in benefitting healthcare. Traditional and conventional diagnostic methods for the infectious disease include a sequence of symptom-based diagnoses followed by extensive identification of pathogens [3]. This consists of several pathological tests like blood, urine, sputum, and imaging techniques like ultrasounds, X-ray, computed tomography (CTs), and magnetic resonance images (MRIs). A patient needs to undergo multiple tests that help the doctor identify the causal organism. These methods are expensive, time-consuming, less sensitive, and labor extensive (require qualitative staff) as well. Machine learning (ML), in general, is a subset of AI that can learn from data and identify patterns without being programmed. It is accomplished by analyzing the existing data and making predictions based on what it has learned from previous experiences. It draws on a range of other disciplines, such as mathematics, statistics, and computer science. ML algorithms have been successfully applied along with the disciplines in several industries, including agriculture, healthcare, marketing, and finance [4]-[6]. Adopting AI techniques in detecting and diagnosing diseases outperforms the conventional approach to diagnosis [7]-[16]. With the latest enhancement in technologies, we can analyze the vast amount of health records and huge databases of images or genomic databases to detect diseases. AI-based diagnosis for infectious disease collects patients' symptoms, medical history, and profile. Then, a model is developed by implementing the ML algorithm, wherein: i) the descriptions of previously solved instances, i.e., data with the correct diagnosis, are loaded and ii) automatically derive medical diagnostic information for new examples. The derived model supports physicians in diagnosing new patients to enhance diagnostic speed, accuracy, and reliability or teaches students or non-specialist physicians how to diagnose patients with a specific diagnostic difficulty. This helps identify the hidden patterns that may contain secrets about the imminence of disease that we never knew [17]. This article highlights how ML algorithms can be used to diagnose the population with infectious diseases using primary clinical data, symptoms, and patient demographic details. Our main objective is to analyze various predictive models considering relevant features for the early diagnosis.

2. MACHINE LEARNING TECHNIQUES

ML is a field of AI that has been around for decades. The field has grown exponentially in the past few years, with new techniques developed daily. ML health therapies are divided into four areas: i) patient diagnostics; ii) patient morbidity or mortality risk assessment, iii) infectious disease outbreaks: prediction and monitoring, and iv) health management planning. The advanced techniques of AI make the diagnosis process more accurate and reliable, which improves the accuracy of diagnosis. ML techniques have become an integral and essential part of the field of diagnosis due to the ability of disease detection.

Technology advancement has incremented the available health data exponentially. Data used in the system is stored and accessed securely [18]. The big data community is a social system where large amounts of healthcare data are shared and processed collaboratively by community members. The objective of the community is to improve the quality and efficiency of healthcare by making use of the latest data-processing and analysis tools. The healthcare industry is one of the most data-rich industries in the world. In today's scenario, healthcare providers and consumers contribute an overwhelming amount of data. The significant way of data collection, organization, and analysis significantly improve healthcare delivery, and considerable significant challenges of electronic health records (EHR) data models are missing values, reasons behind these missing values, data model validity, the need for various types of data, and operational feasibilities [19]. Pandey and Janghel [17] have discussed different ML techniques for EHR data predicting diseases' onset. Cruz and Wishart [7], Chandru and Seetharam [20] highlighted the practice of testing a model with multiple machine-learning techniques and using techniques such as robust feature selection and adequate data size to improve models for routine clinical procedures and hospital settings. This study emphasizes that various ML techniques have been effective in predicting infectious diseases. We reviewed several ML models for tuberculosis, malaria, flu, dengue, COVID-19, cystitis, nonspecific urethritis, and other diseases.

2.1. Logistic regression

Logistic regression (LR) is a statistical technique used to model the relationship between one or more categorical explanatory variables and a continuous response variable. Sigmoid function converts the independent variable (X) into an expression of probability between 0 and 1 concerning the dependent variable (Y).

$$y = \frac{1}{1 + e^{-x}}$$

Where: $e = Euler's\ constant$

Ruano-ravina *et al.* [21] have been used to analyze the relationship between cigarette smoking and death from lung cancer.

2.2. K-nearest neighbor

K-nearest neighbor (KNN) is a supervised ML algorithm capable of performing both classifications and regression tasks using numbers (K) of neighbors (instances). It finds the k most similar observations to each query point and assigns the average of their outcomes as the predicted value [22]. KNN algorithm is used in developing different disease diagnosis models [23].

2.3. Support vector machine

Support vector machine (SVM) learns a representation of all data points such that separate labels/categories are divided or separated by a clear gap, which is as large as possible. The algorithm finds the optimal hyperplane that separates the clusters/classes. The vectors near the hyperplane support vectors [24]. Let training set $\{(xi, yi)\}\ i = 1..n, xi \in Rd, yi \in \{-1, 1\}$ be separated by a Hyperplane with margin ρ then distance from example x_i to the separator is,

$$r = \frac{w^T x_i + b}{\|w\|}$$

commonly used SVM algorithms are the support vector regression, least squares SVM, and successive projection algorithm-SVM [4]. SVMs are widely used in pattern recognition and classification and have been effectively used in various real-world problems [25]-[27].

2.4. Decision tree

A decision tree (DT) displays all possible scenarios (i.e., decisions) and outcomes (i.e., results) by: i) splitting the dataset into two groups, either based on the value of an attribute or randomly and then, ii) comparing the two groups to identify the best split point, iii) step i) and ii) is done until all of the data points have been categorized. Few algorithms to build a DT are classification and regression trees (CART), iterative dichotomiser (ID), and C4.5, chi-squared automatic interaction detection (CHAID) [28]. DTs are used to make decisions, classify objects, predict outcomes, and analyze data [29]. DTs are employed in medicine for disease diagnosing and drug discovery [30].

2.5. Naive Bayes

A Naive Bayes (NB) classifier is a probabilistic classifier that uses Bayes' theorem with some simplifications as its foundation. The label, it is assumed that the features are conditionally independent in NB classifier [31]. NB has been applied to medical diagnosis, spam-filtering, and weather forecasting [6], [32].

2.6. Neural networks

Neural network (NN) algorithms are inspired by the structure of the human brain and are made up of layers of neurons. These neurons are arranged in layers: input layers, hidden layers, and output layers. These channels are assigned with weight and bias (some numerical value). The summation of the product of inputs and corresponding weights are calculated:

$$\text{sum} = \sum_{i=1}^{n} x_i w_i + b$$

this computed value is sent as input to the neurons in the hidden layers. The calculated value is then passed to the threshold function to activate a neuron.

$$\text{y} = f(\sum_{i=1}^{n} x_i w_i + b)$$

Few standard NN are artificial (ANN), convolutional (CNN), and recurrent (RNN). In ANN [33], a series of neurons are interconnected by feed-forward and back-propagation forms. CNN uses a variation of multi-layer perceptron [34]. ANN models have huge hardware dependency. However, the CNN model does not encode the object's position and orientation and needs a lot of training data to work efficiently. These networks can be trained to recognize visual patterns, including voice, handwriting, and image recognition [27], [35]-[40]. As a result, modern NNs are significantly more potent than their predecessors.

2.7. Ensemble techniques

Ensemble in ML techniques combines several base models to provide one optimal predictive model/learners that help to improve the result.

a. Bagging

It creates several copies of the training data and then trains a separate classifier on each copy. The predictions made by these classifiers are then combined to produce a final prediction.

b. Boosting

It is a sequential algorithm that works by constructing a series of homogenous weak classifiers, each of which is designed to correct the errors made by the previous classifier. The first model is trained on the entire dataset, and the second model is on the first model’s predictions. The final classifier is the combination of all the weak classifiers.

c. Stacking

Stacking ensemble algorithm functions by constructing heterogeneous weak classifiers (base model) and using a meta-model. This meta-model learns to combine predictions of base models. In stacking, the combining mechanism is that the output of the classifiers (level 0 classifiers) will be used as training data for another classifier (level 1 classifier) to approximate the same target function.

A random forest is a form of ensemble learning method comprising several decision trees, i.e., multiple classifiers. It is used for solving both classification and regression problems. Steps to build a random forest are: i) take input variables, ii) randomly select subsets of input variables as candidate inputs for each tree, iii) builds a tree with selected candidate predictors, iv) find the prediction from all of the trees in the forest, and v) compute the average of all predictions. It is utilized effectively when we need to predict an instance's class based on certain characteristics. A random forest can handle: numerical and categorical data, although it works best with continuous response variables. It's frequently combined with other methods like logistic regression, linear regression, and support vector machine [41].

3. METHOD

For this review, we performed searches on online databases, including ScienceDirect, Scopus, PubMed, and Google Scholar. The following inclusion criteria were used: keywords, such as infectious diseases, tuberculosis (TB), influenza, human immunodeficiency virus (HIV), COVID-19, malaria, dengue, pneumonia, urinary tract infections (UTI), and bacterial colonies infection. In the document type, we included journal paper, conference proceedings, book, book chapter; search related to thematic areas, we included AI, ML, and deep learning.

The following exclusion criteria were also used: not related to the diagnosis of diseases. Not belonging to the domain of ML, deep learning, or AI; not covering the year of publication between 2018 and

2022; not in the English language; not conclusive; not relevant. The initial search yielded 1,056 items. After applying inclusion criteria, 27 papers were selected for our studies.

4. RESULTS AND DISCUSSION

Data-driven approaches and decision-making systems are the two key points when ML and healthcare are combined for effective disease prediction and diagnosis. It is observed that 70.58% of research articles used SVM, 29.41% used ANN, 41.17% used random forest, 35.29% employed KNN and LR techniques, 23.52% used NB, DT and "boosting" ensemble method, and 5.88% used CNN techniques for infectious disease diagnostic models, as shown in Figure 1. Others ML techniques include J48, adaptive network based fuzzy inference system (ANFIS), long short-term memory (LSTM), and vector quantization (VQ).

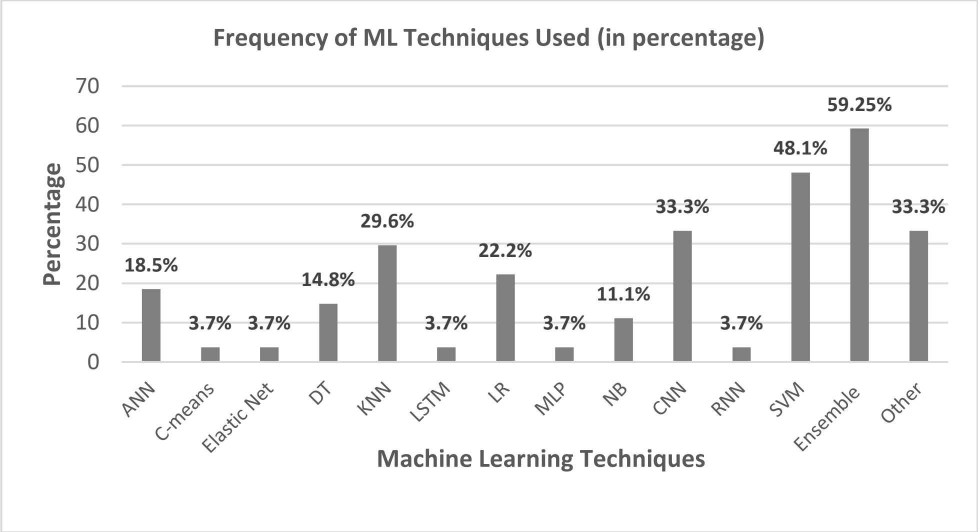

Figure 1. Distribution of ML techniques used in survey papers (in %)

Fuhad *et al*. [42] have proposed an automatic model for malaria detection in blood smears. The author used medical images in his study and obtained an accuracy of 99.5 percent on 28×28 images using the Autoencoder-based training method. The proposed model is computationally impressive because it only requires 4,600 flops, whereas the previous model took over 19.6 billion flops. Tuberculosis is one of the world's leading causes of death. In 2016, there were 10.4 million new cases and 1.7 million deaths from tuberculosis. TB is an infectious disease caused by bacteria. It usually affects the lungs but can also affect other body parts. In the absence of proper treatment, this leads to a fatal disease. Transcription of patient sputum samples used for TB diagnosis is critical and time-consuming. The test results help determine the best course of treatment for the patient. There are two major limitations of the current process: time and cost of diagnosis. Osmor and Okezie [43] have proposed an efficient model for TB diagnosis using transcriptional signatures obtained from a patient blood sample. The model employs SVM and NB techniques to get higher accuracy using the weighted ensemble technique.

The flu, also known as influenza, is a severe viral infection, and the diagnosis is complicated. Clinical diagnosis can be difficult due to the similarity of symptoms with other respiratory illnesses like asthma, pneumonia, and other viral diseases such as the common cold. Based on the patient's signs and symptoms, Marquez and Barron [44] have developed a model for intelligent influenza diagnosis. Their study employed a dataset of 3,346 samples from Mexico's National System of Epidemiological Surveillance (SINAVE) with 1,484 controls and 1,862 cases. To partition the total samples into training and testing sets, the author used a 5-fold cross-validation procedure. The findings reveal that SVM outperforms other models in terms of accuracy (0.9524), sensitivity (0.9715), and specificity (0.9285). Other ML techniques are multilayer perceptron (MLP), C-means, and VQ. However, the author found it difficult to establish the appropriate architecture while implementing the MLP technique, and the performance is slow. They emphasized the importance of working with multiple ML models with fewer signs and symptoms.

The COVID-19 pandemic has brought unprecedented challenges to businesses and individuals across the globe. ML has been used to predict outbreaks, develop therapies, and create simulations to help decision-makers. Zoabi *et al.* [45] have presented a ML-based (gradient boosting) model for predicting COVID-19 diagnosis. The data was collected from the Israeli Ministry of Health. The proposed model works based on eight binary features: sex, age > 60 years, known contact with an infected individual and five basic clinical symptoms. The clinical symptoms include cough, fever, sore throat, shortness of breath, and headache. Yousif *et al.* [46] focused on the COVID-19 disease diagnostic system that combines the ML classification models with specific implementations of the LR, SVM, and extreme gradient boosting (XGBoost) techniques. The author has collected 300 samples from private laboratories in Iraq or Baghdad, where 87 cases are infected. XGBoost classier surpasses the competition with an overall accuracy of 0.87 and an F1 score of 0.91. However, the author suggested that the deep learning techniques on larger datasets potentially improve the model's overall performance. Muhammad *et al.* [47], have focused on COVID-19 prediction using a classification model in ML with the specific implementation of decision trees, logistic regression, NB, support vector machines, and artificial neural networks. They include a screening and epidemiology dataset of positive and negative COVID-19 cases in Mexico City with a sample size of 263,007 and 41 features. Before creating the model, the correlation coefficient analysis was performed on the demographic and clinical reverse-transcriptase polymerase chain reaction (RT-PCR) features. The demographic details include age, sex, pneumonia, diabetes, asthma, hypertension, obesity, cardiovascular diseases (CVDs), chronic kidney diseases (CKDs), and tobacco. The accuracy of the decision tree model is 94.99%. This reveals that males who smoke tobacco with age above 45 are more susceptible to SARS-CoV-2. The author obtained specificity of 93.34% and 94.3% for the SVM and NB models, respectively. Ensemble technique became state of art that produces better results than the existing models [48], [49]. Table 2 (see in Appendix) [23], [27], [32], [35]-[47], [50]-[61] summarizes reviewed articles on infectious disease diagnosis using various ML techniques. As ML techniques have shown their potential in medical systems, there is still sufficient potential to rise in various areas [14], [62].

4.1. Data types

Electronic health records (EHR) contain both structured and unstructured data types. Unstructured data includes various clinical notes, reports, discharge summaries, images, audio, and videos of patients. Structured data alone do not provide all of the information associated with clinical context. Unstructured data often provide additional, valuable information. However, utilizing these data involves complex and time-consuming analytic operations and requires many manual efforts. Chen *et al.* [63] to predict cerebral infarction disease, the author created a CNN-based multimodal risk prediction algorithm. Three years of real-world hospital data (2013-2015) were collected from the hospital, including patient demographic details, patient narration of illness, doctor's interrogation records, and medical history. The author also created three datasets- i) S-data: which contains only patient structure data, ii) T-data includes only textual data, and iii) S&T-data includes both S-data and T-data. They further developed a CNN-based unimodal disease risk prediction model for T-data and a CNN-based multimodal disease risk prediction model for S&T-data, as well as NB, KNN, DT, ML algorithms for the prediction of cerebral infarction on S-data. The author observed that cerebral infarction disease could be predicted with up to 94.80% accuracy based on the proposed CNN-based multimodal disease risk prediction model.

4.2. Data volumes

There are enormous challenges in detecting and diagnosing multiple infectious diseases using ML techniques. The rapid and exponential increase in the data has found a challenge regarding prediction accuracy. Various ML algorithms help find the hidden pattern based on the patient's symptoms [18]. ML technique seeks attention from researchers, such as the adoption of unsupervised (clustering) and deep learning (neural network) model [8]. The rapid grow able nature of data and the requirement of maintaining the accuracy of detection and diagnosis of infectious diseases need such a hybrid system that can handle both the conditions.

4.3. Data quality and temporality

Data temporality is the study of data that changes over time. This is a severe problem in disease diagnosis because each patient's data could have different timeframes and the quality of each dataset varies. For example, in the case of breast cancer, mammography screening changes over time for each woman. Studies show that the accuracy of mammography screening changes month-to-month based on age, mammogram use history, and even breast density. The rate at which data is collected, processed, and made available to the public varies by organization. This can lead to errors and disputes in diagnostic results because there are no checks on the quality of the data [52].

4.4. Small size of dataset, sample size or validation-set

In the context of ML, the "curse of dimensionality" is the issue of holding too many features and too fewer samples. One of the challenges in building ML models for infectious diseases is very few larger size samples available. Discovering patterns from a small dataset that is representative of the whole often leads to biased results [52]. There are only a few cases in the entire world, so there is not enough data to accurately predict when someone gets an infection. The size of the dataset may be small because there are only a few samples, or a lot of data is unavailable in the dataset. Including a smaller sample size produces a bias towards more severe infection. This does not provide a comprehensive and accurate analysis of the data. For example, for HIV [56], about 80 million people have been infected and are alive today. However, the data on their progression to acquired immunodeficiency syndrome (AIDS) and death is not available. Therefore, any ML model trained on this data would be biased towards predicting AIDS in general and not just HIV. This is because of the limited information available in the dataset. The researchers needed to figure out how to get enough varied data into their database to build up a good initial representation or at least one that mirrors the real world as closely as possible.

4.5. Lack of models to deal directly with real-world data

The most significant ML application is the diagnosis of infectious diseases. However, a key constraint in this field is the lack of models that deal directly with real-world data. The traditional modeling methods are not suitable for data with missing values, irregular spacing, and high-dimensional features. The majority of the current ML models are developed and evaluated either on artificially generated data or are experimental [11], [22], [43], [45], [55], [56]. In many cases, these models perform better than existing methods based on evaluation metrics such as accuracy and precision. However, the generalization of these models is often unknown as they have not been evaluated on datasets corresponding to the target disease.

4.6. Cost and time of diagnosis

Infectious disease diagnosis is a challenging task for ML. There are several reasons for this: i) many infectious diseases have similar symptoms, making it difficult to differentiate between them, ii) the tests required to diagnose infection are expensive and time-consuming [43], iii) the data set used to train a ML algorithm is usually incomplete or inaccurate, and iv) the algorithm cannot account for all of the factors that contribute to a diagnosis. Disease diagnosis is one of the most expensive processes in healthcare. Even after a diagnosis has been made, finding a cure or appropriate treatment options takes years. This delays healing prolongs illness, and often leads to disability or death. A more pragmatic approach is needed to improve efficiency and reduce costs while tailoring treatments to individual patients based on their genetic makeup and personal medical history. The time and cost required for the diagnosis will depend on various factors, including the complexity of the disease, the accuracy and specificity of the ML algorithm, and the availability of data. Early detection and automated diagnosis have become important and necessary requirements.

5. CONCLUSION

Infectious diseases account for about one-sixth of all deaths worldwide and cause immense human suffering every year. It also imposes a significant economic burden. For example, in the United States, the centers for disease control and prevention (CDC) estimate that the overall cost of infectious diseases is more than $120 billion annually. This includes physician and clinical expenditures, prescription drug expenditures, and hospital expenditures. This is the biggest challenge in low-income countries where medical services are scarce. It takes days to get the results from a traditional lab test and costs money which many people do not have access to. The World Health Organization estimates only one infectious disease doctor per 100,000 in low-income country populations. The study's primary focus is on the use of datasets that exploit some primary clinical data, symptoms, and profiles of patients. The datasets are further analyzed using ML techniques for the early diagnosis of disease. The study found limitations in data handling issues. These include data types, volume, quality, temporality, and availability. ML provides several solutions to the problem of time and costs in diagnosing infectious diseases. Automated diagnostics can provide accurate results faster with a better diagnosis than what is possible with human labor. There is a need to explore access to medical care in developing countries. In this study, we review the recent advances in ML algorithms and focus on their potential for diagnosing infectious diseases. The study found that supervised ML techniques are widely used for diagnosis. To improve the overall performance of predictive model, ensemble techniques are being employed instead of a traditional ML classifier.

APPENDIX

Table 2. Diagnosis of infectious diseases

Paper/Ref.	Disease	ML techniques	Types of data	Sample size	Performance
[23]	UTI	KNN	Structured	NA	97.4% accuracy when applying the suggested value of k=6
[27]	Bacterial colonies infection	Combination of Deep learning models (ResNet 101 CNN architecture) & SVM	Unstructured (Image)	44,985	99.61% accuracy, 99.58% recall, 99.58% precision, and 99.97% specificity
[32]	Dengue Hemorrhagic Fever	NB, SVM, RF	Structured	213	NB: AUC=0.715, CA=0.698, F1 Score=0.743, Precision=0.775, Recall=0.71; SVM: AUC=0.512, CA=0.445, F1 Score=0.488, Precision=0.560, Recall=0.433; RF: AUC=0.898, CA=0.796, F1 Score=0.831, Precision=0.811, Recall=0.822
[35]	COVID–19, Pneumonia	CNN	Unstructured (Image)	3,788	For multiclass classification (COVID-19, pneumonia, and normal), Accuracy=97.9% (loss of 0.052) and for binary classification (COVID-19 and normal), Accuracy=99.8%, sensitivity= 99.52%, specificity=100% (loss of 0.001)
[36]	Viral Pneumonia	CNN	Unstructured (Image)	744	Training accuracy=91% and Training loss=0.63, Validation accuracy=81% and Validation loss=0.7108
[37]	COVID–19	CNN	Unstructured (Image)	2,905	accuracy of 97.44% and training accuracy of 97.55%.
[38]	COVID-19	CNN	Unstructured (Image)	2,000	accuracy, sensitivity, specificity, F1-score, and area under curve (AUC) of 0.98, 0.97, 0.98, 0.97, and 0.99 respectively
[39]	COVID-19	CNN	Unstructured (Image)	188	98.68% accuracy, 100% precision, and 100% specificity. 97.37%, 98.67%, and 98.68% for sensitivity, F-measure, and Gmean, respectively.
[40]	Tuberculosis	Ensemble Technique	Unstructured (Image)	788	accuracy of 93.59%, sensitivity of 92.31% and specificity of 94.87%
[42]	Malaria	Autoencoder, CNN-SVM, CNN-KNN	Unstructured (Image)	13,029	Image Size:(28,28), Autoencoder, F1 score: 0.9951, Precision: 0.9929, Sensitivity: 0.9880, Specificity:0.9917 Image Size:(32,3), Autoencoder, F1 score: 0.9922, Precision: 0.9892, Sensitivity: 0.9952, Specificity:0.9917 Image Size:(32,32), CNN-SVM, F1 score:0.9918, Precision: 0.9921, Sensitivity: 0.9916 Image Size:(32,32), CNN-KNN, F1 score:0.9928, Precision: 0.9911, Sensitivity: 0.9923
[43]	Tuberculosis	SVM, NB, Ensemble Techniques	Structured (Genome)	456	SVM: Accuracy, sensitivity and specificity of 0.92, 0.98, and 0.66 respectively. NB: Accuracy, sensitivity and specificity of 0.87, 0.87 and 0.88 respectively. Ensemble Techniques: Accuracy, sensitivity and specificity of 0.95, 0.94 and 0.95 respectively.
[44]	Influenza	SVM, C-Means, MLP, Vector quantization	Structured	3,346	SVM: Accuracy, sensitivity and specificity of 0.9412, 0.965, and 0.9029 respectively. MLP: Accuracy, sensitivity and specificity of 0.8557, 0.919 and 0.8306 respectively. VQ: Accuracy, sensitivity and specificity of 0.7523, 0.785 and 0.754 respectively. C-means: Accuracy, sensitivity, and specificity of 0.8038, 0.8162 and 0.7831, respectively.
[45]	COVID-19	Gradient boosting	Structured	47,401 patients (3,624 positive)	AUC of 0.862, auPRC (area under the precision-recall curve) of 0.66 with 95%, CI: 0.647–0.678.
[46]	COVID-19	LR, XGBoost, SVM	Structured	300 (87 positive)	LR: Accuracy=0.81 Sensitivity=0.96 Specificity=0.42 F1-Score=0.88. SVM: Accuracy=0.82 Sensitivity=0.98 Specificity=0.42 F1-Score=0.89. XGBoost classifier: Accuracy=0.87 Sensitivity=0.94 Specificity=0.69 F1-Score=0.91

Table 2. Diagnosis of infectious diseases (continue)

Paper/Ref.	Disease	ML Techniques	Types of Data	Sample Size	Performance
[47]	COVID-19	LR, DT, SVM, NB, and ANN	Structured	263,007	DT: Accuracy=94.99%, Sensitivity=89.2%, Specificity=93.22%. LR: Accuracy=94.41, Sensitivity=86.34, Specificity=87.34. NB: Accuracy=94.36, Sensitivity=83.76, Specificity=94.3. SVM: Accuracy=92.4, Sensitivity=93.34, Specificity=76.5. ANN: Accuracy=89.2, Sensitivity=92.4, Specificity= 83.3
[50]	COVID-19	XGBoost	Structured	413 patients	Sensitivity of 92.5% and specificity of 97.9%
[51]	COVID-19	LR	Structured	43,752 surveys (498 self-reported COVID-19 positive)	AUC of 0.737
[52]	COVID-19	ANN, CNN, RNN, CNN-LSTM, and CNNRNN, and (LSTM selected)	Structured	600 patients (80 positive)	AUC of 62.50%, the accuracy of 86.66%, precision of 86.75%, recall of 99.42%, F1-score of 91.89%
[53]	COVID-19	J48, ANFIS (adaptive neuro-fuzzy inference system), KNN, SVM, ANN, RF, LR and gradient boosting	Structured	260	LR: observed as best model in terms of accuracy and F-measure by 1.4765% and 1.2782, respectively
[54]	Urinary Tract Infection (UTI)	DT, SVM, RF, ANN	Structured	59	DT: Accuracy=93.22%, Sensitivity=95.55%, specificity=85.71%. SVM: Accuracy=96.61%, Sensitivity=97.77%, specificity=92.85. RF: Accuracy=96.61%, Sensitivity=95.55%, specificity=100%. ANN: Accuracy=98.30$, Sensitivity=97.77%, specificity=100%"
[55]	Tuberculosis	LR, SVM, KNN, DT, RF, NN	Structured	113	RF in terms of accuracy 0.7447 and specificity=0.7699. SVM in terms of Sensitivity=0.809
[56]	HIV	Elastic Net, KNN, RF, SVM, XGBoost, and light gradient boosting (LGBT) algorithms.	Structured	80,000	XGBoost by F1 Score of 90% for gender = male and 92% for gender = female
[57]	Influenza	SVM, KNN, RF, ANN	Structured	9,548	RF in terms of accuracy 0.869 and specificity=0.9277; SVM in terms of Sensitivity=0.8211
[58]	COVID-19	RF, DT, SVM, KNN, and LR	Structured	10,000	RF, DT, SVM, KNN, and LR with an accuracy of (0.88%, 0.88%, 0.87, 0.86, and 0.88%) respectively
[59]	Tuberculosis	Ensemble Technique	Unstructured (Image)	788	accuracy of 89.77%, sensitivity of 90.91% and specificity of 88.64%.
[60]	COVID-19	KNN	Unstructured (Image)	746	For the combination of Haralick and local binary pattern feature extraction, Accuracy of 93.30% and For the combination of Haralick, histogram, and local binary pattern the best area under the curve (AUC) = 0.948. Proposed models outperform CNN by a 4.3% margin.
[61]	COVID-19	Transfer learning approach with CNN models (inception-V3, the Xception, and the MobileNet)"	Unstructured (Image)	7,800	F1-score is 100% in the first task and 97.66 in the second task

REFERENCES

[1] F. Balloux and L. V. Dorp, “Q&A: What are pathogens, and what have they done to and for us?,” *BMC Biology*, vol. 15, no. 1, pp. 1–6, 2017, doi: 10.1186/s12915-017-0433-z.

[2] J. M. V. Seventer and N. S. Hochberg, “Principles of infectious diseases: transmission, diagnosis, prevention, and control,” *International encyclopedia of public health*, p. 22, 2017, doi: 10.1016/B978-0-12-803678-5.00516-6.

[3] S. Agrebi and A. Larbi, “Use of artificial intelligence in infectious diseases,” in *Artificial Intelligence in Precision Health*, Elsevier, 2020, pp. 415–438. doi: 10.1016/b978-0-12-817133-2.00018-5.

[4] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018, doi: 10.3390/s18082674.

[5] A. Callahan and N. H. Shah, "Machine learning in healthcare," in *Key Advances in Clinical Informatics: Transforming Health Care through Health Information Technology*, 2017, pp. 279–291, doi: 10.1016/B978-0-12-809523-2.00019-4.

[6] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, p. e01802, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[7] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006, doi: 10.1177/117693510600200030.

[8] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," *Journal of personalized medicine*, vol. 10, no. 2, p. 21, 2020, doi: 10.3390/jpm10020021.

[9] N. Alballa and I. A. -Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review," *Informatics in Medicine Unlocked*, vol. 24, p. 100564, 2021, doi: 10.1016/j.imu.2021.100564.

[10] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, Jan. 2015, doi: 10.1016/j.csbj.2014.11.005.

[11] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Applied Intelligence*, vol. 52, no. 3, pp. 2411–2422, 2022, doi: 10.1007/s10489-021-02533-w.

[12] S. A. Taie and W. Ghonaim, "A new model for early diagnosis of alzheimer's disease based on bat-svm classifier," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 759–766, Apr. 2021, doi: 10.11591/eei.v10i2.2714.

[13] O. Hrizi *et al.*, "Tuberculosis disease diagnosis based on an optimized machine learning model," *Journal of Healthcare Engineering*, vol. 2022, 2022, doi: 10.1155/2022/8950243.

[14] M. D. Pholo, Y. Hamam, A. B. Khalaf, and C. Du, "Classifying lymphoma and tuberculosis case reports using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2857–2865, Oct. 2021, doi: 10.11591/eei.v10i5.3132.

[15] S. H. Abdulla, A. M. Sagheer, and H. Veisi, "Breast cancer segmentation using K-means clustering and optimized region-growing technique," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 158–167, Feb. 2022, doi: 10.11591/eei.v11i1.3458.

[16] Z. Saringat, A. Mustapha, R. D. R. Saedudin, and N. A. Samsudin, "Comparative analysis of classification algorithms for chronic kidney disease diagnosis," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1496–1501, Dec. 2019, doi: 10.11591/eei.v8i4.1621.

[17] S. K. Pandey and R. R. Janghel, "Recent deep learning techniques, challenges and its applications for medical healthcare system: a review," *Neural Processing Letters*, vol. 50, no. 2, pp. 1907–1935, 2019, doi: 10.1007/s11063-018-09976-2.

[18] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," in *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, Mar. 2019, pp. 1211–1215. doi: 10.1109/ICCMC.2019.8819782.

[19] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial intelligence and human trust in healthcare: focus on clinicians," *Journal of medical Internet research*, vol. 22, no. 6, p. e15154, 2020, doi: 10.2196/15154.

[20] A. S. Chandru and K. Seetharam, "Framework for efficient transformation for complex medical data for improving analytical capability," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, pp. 4853–4862, Oct. 2020, doi: 10.11591/ijece.v10i5.pp4853-4862.

[21] J. Ruano-Ravina, A. Figueiras, A. Montes-Martinez, and A. Barros-Dios, "Dose–response relationship between tobacco and lung cancer : new findings," *European journal of cancer prevention*, vol. 12, no. 4, pp. 257–263, 2003, doi: 10.1097/00008469-200308000-00003.

[22] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.

[23] M. N. F. Jamaluddin *et al.*, "An Application of presumptive diagnosis for urinary tract infection via kNN algorithm approach," *Charting the Sustainable Future of ASEAN in Science and Technology*, 2020, pp. 377–388. doi: 10.1007/978-981-15-3434-8_32.

[24] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12. pp. 1565–1567, 2006, doi: 10.1038/nbt1206-1565.

[25] V. Jakkula, "Tutorial on support vector machine (SVM)," *School of EECS, Washington State University*, vol. 37, no. 2.5, pp. 1–13, 2011, [Online]. Available: http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf.

[26] L. Auria and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis," *SSRN Electron. J.*, 2008, doi: 10.2139/ssrn.1424949.

[27] S. A. Akbar, K. H. Ghazali, H. Hasan, Z. Mohamed, W. S. Aji, and A. Yudhana, "Rapid bacterial colony classification using deep learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 6, no. 1, pp. 352–361, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp352-361.

[28] S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data mining and knowledge discovery*, vol. 2, no. 4, pp. 345–389, 1998, doi: 10.1023/A:1009744630224.

[29] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004, doi: 10.1002/cem.873.

[30] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and prediction of diabetes disease using decision tree method," in *CEUR Workshop Proceedings*, vol. 2824, pp. 163–172, 2021.

[31] M. N. Murty and V. S. Devi, *Pattern Recognition*, 1st ed. Springer London, 2011. doi: 10.1007/978-0-85729-495-1.

[32] R. Arafiyah, F. Hermin, I. R. Kartika, A. Alimuddin, and I. Saraswati, "Classification of dengue haemorrhagic fever (DHF) using SVM, naive bayes and random forest," in *IOP Conference Series: Materials Science and Engineering*, vol. 434, no. 1, p. 012070, 2018, doi: 10.1088/1757-899X/434/1/012070.

[33] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *Journal of microbiological methods*, vol. 43, no. 1, pp. 3–31, 2000, doi: 10.1016/S0167-7012(00)00201-3.

[34] T. Liu, S. Fang, Y. Zhao, P. Wang, and J. Zhang, "Implementation of training convolutional neural networks," *arXiv Prepr. arXiv1506.01195*, 2015, doi: 10.48550/arXiv.1506.01195.

[35] A. Musha, A. Al Mamun, A. Tahabilder, J. Hossen, B. Jahan, and S. Ranjbari, "A deep learning approach for COVID-19 and pneumonia detection from chest X-ray images," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12,

no. 4, pp. 3655–3664, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3655-3664.
[36] B. N. Chaithanya, T. J. S. Jain, A. U. Ruby, and A. Parveen, "An approach to categorize chest X-ray images using sparse categorical cross entropy," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1700–1710, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1700-1710.
[37] M. Masadeh, A. Masadeh, O. Alshorman, F. H. Khasawneh, and M. A. Masadeh, "An efficient machine learning-based COVID-19 identification utilizing chest X-ray images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 356–366, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp356-366.
[38] L. N. Mahdy, A. I. B. El Seddawy, and K. A. Ezzat, "Automatic COVID-19 lung images classification system based on convolution neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 5573–5579, Oct. 2022, doi: 10.11591/ijece.v12i5.pp5573-5579.
[39] E. H. Ahmed, M. R. M. Alsemawi, M. H. Mutar, H. O. Hanoosh, and A. H. Abbas, "Convolutional neural network for the detection of coronavirus based on X-ray images," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 37–45, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp37-45.
[40] S. K. T. Hwa, A. Bade, M. H. A. Hijazi, and M. S. Jeffree, "Tuberculosis detection using deep learning and contrast-enhanced canny edge detected X-Ray images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 4, pp. 713–720, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp713-720.
[41] D. Sarkar, R. Bali, and T. Sharma, *Practical Machine Learning with Python*, 1st ed. Apress Berkeley, CA, 2018. doi: 10.1007/978-1-4842-3207-1.
[42] K. M. F. Fuhad, J. F. Tuba, M. R. A. Sarker, S. Momen, N. Mohammed, and T. Rahman, "Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application," *Diagnostics*, vol. 10, no. 5, p. 329, 2020, doi: 10.3390/diagnostics10050329.
[43] V. C. Osamor and A. F. Okezie, "Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-94347-6.
[44] E. Marquez and V. Barron, "Artificial intelligence system to support the clinical decision for influenza," *2019 IEEE International Autumn Meeting on Power, Electronics and Computing, ROPEC*, 2019, pp. 1–5. doi: 10.1109/ROPEC48299.2019.9057056.
[45] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj digital medicine*, vol. 4, no. 1, pp. 1–5, 2021, doi: 10.1038/s41746-020-00372-6.
[46] A. Y. Yousif, S. M. Younis, S. A. Hussein, and N. M. G. Al-Saidi, "An intelligent computing for diagnosing covid-19 using available blood tests," *International Journal of Innovative Computing, Information and Control*, vol. 18, no. 1, pp. 57–72, 2022, doi: 10.24507/ijicic.18.01.57.
[47] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset," *S SN computer science*, vol. 2, no. 1, pp. 1–13, 2021, doi: 10.1007/s42979-020-00394-7.
[48] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.
[49] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: a comprehensive review," *Healthcare*, vol. 10, no. 3. p. 541, Mar. 2022. doi: 10.3390/healthcare10030541.
[50] W. T. Li *et al.*, "Using machine learning of clinical data to diagnose COVID-19," *BMC medical informatics and decision making*, vol. 20, no. 1, p. 247, 2020, doi: 10.1186/s12911-020-01266-z.
[51] S. Shoer *et al.*, "A prediction model to prioritize individuals for a SARS-CoV-2 test built from national symptom surveys," *Med*, vol. 2, no. 2, pp. 196–208, 2021, doi: 10.1016/j.medj.2020.10.002.
[52] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos, Solitons and Fractals*, vol. 140, p. 110120, 2020, doi: 10.1016/j.chaos.2020.110120.
[53] N. Kumar, N. N. Das, D. Gupta, K. Gupta, and J. Bindra, "Efficient automated disease diagnosis using machine learning models," *Journal of Healthcare Engineering*, vol. 2021, 2021, doi: 10.1155/2021/9983652.
[54] I. A. Ozkan, M. Koklu, and I. U. Sert, "Diagnosis of urinary tract infection based on artificial intelligence methods," *Computer methods and programs in biomedicine*, vol. 166, pp. 51–59, 2018, doi: 10.1016/j.cmpb.2018.10.007.
[55] M. H. Ali, D. M. Khan, K. Jamal, Z. Ahmad, S. Manzoor, and Z. Khan, "Prediction of Multidrug-resistant tuberculosis using machine learning algorithms in SWAT, Pakistan," *Journal of healthcare engineering*, vol. 2021, 2021, doi: 10.1155/2021/2567080.
[56] C. K. Mutai, P. E. McSharry, I. Ngaruye, and E. Musabanganji, "Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa," *BMC medical research methodology*, vol. 21, no. 1, pp. 1–11, 2021, doi: 10.1186/s12874-021-01346-2.
[57] S. Maheshwari, A. Sharma, R. Kumar, and Pratyush, "Early detection of influenza using machine learning techniques," in *Recent Innovations in Computing*, Springer, Singapore, 2022, pp. 111–124. doi: 10.1007/978-981-16-8892-8_9.
[58] A. H. Ahmed, M. N. A. Al-Hamadani, and I. A. Satam, "Prediction of COVID-19 disease severity using machine learning techniques," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1069–1074, Apr. 2022, doi: 10.11591/eei.v11i2.3272.
[59] S. K. T. Hwa, M. H. A. Hijazi, A. Bade, R. Yaakob, and M. S. Jeffree, "Ensemble deep learning for tuberculosis detection using chest X-Ray and canny edge detected images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 429–435, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp429-435.
[60] R. A. Nugroho, A. S. Nugraha, A. Al Rasyid, and F. W. Rahayu, "Improvement on KNN using genetic algorithm and combined feature extraction to identify COVID-19 sufferers based on CT scan image," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 5, pp. 1581–1587, Oct. 2021, doi: 10.12928/TELKOMNIKA.v19i5.18535.
[61] M. Al-Smadi, M. Hammad, Q. B. Baker, S. K. Tawalbeh, and S. A. Al-Zboon, "Transfer deep learning approach for detecting coronavirus disease in X-ray images," *International Journal of Electrical and Computer Engineering (IJECE)* , vol. 11, no. 6, pp. 4999–5008, Dec. 2021, doi: 10.11591/ijece.v11i6.pp4999-5008.
[62] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017, doi: 10.1093/bib/bbx044.
[63] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," in *IEEE Access*, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

BIOGRAPHIES OF AUTHORS

Smriti Mishra is pursuing Ph.D. in Computer Science from Magadh University, Bodhgaya, India. She obtained master of computer application (MCA) degree from Jawaharlal Technological University, India 2011. She received bachelors of computer application (BCA) from Magadh University, Bodh Gaya in 2008. Her research interest includes data science, machine learning, image processing, and artificial intelligence. She can be contacted at email: smritiit.gaya@gmail.com.

Ranjan Kumar is an associate professor at Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India. He received his master degree (MCA) from Department of Computer Science, University of Delhi in 2003 and completed his Ph.D. in software reliability in the year 2019. His research interests include software reliability, machine learning, data science, and bio medical computation. He can be contacted at email: ranjan301@gmail.com.

Sanjay Kumar Tiwari is currently working as associate professor at P.G. Department of Mathematics, Bodh Gaya, India. He has received his Ph.D. from Magadh University, Bodh Gaya, India. He has more than 20 years of academic experience. His area of interest includes fixed-point theorem, real analysis, software engineering, and data science. He is contributing as a reviewer of various reputed journals. He can be contacted at email: tiwari.dr.sanjay@gmail.com.

Priya Ranjan graduated from IIT Kharagpur (EE, 1997), India and earned M.S. (EE) and Ph.D. (ECE) degrees from the University of Maryland, College Park, USA in 1999 and 2003, respectively. To pursue his passion in discovering new ideas in control theory and applications, he joined the Center for Artificial Intelligence and Robotics (CAIR, Bangalore), a DRDO lab under the tutelage of Prof. M. Vidyasagar. To continue and expand his studies, in 1997 he moved to the Computer Aided Control Systems Engineering (CACSE) lab at the Institute for Systems Research (ISR) at the University of Maryland under the guidance of Prof. E. H. Abed where he would spend the next twelve years in Maryland in various capacities. His latest interest is in biological computation and oncological image processing. He has received many large project awards from some of the most reputed agencies like National Science Foundation (NSF)-USA, DARPA (credited with leading development of the Internet)-USA, Delhi University, DBT, DST-RFBR, and ICMR. He can be contacted at pranjan@gmail.com.