REVIEW ARTICLE

Computer Based Diagnosis of Some Chronic Diseases: A Medical Journey of the Last Two Decades

Samir Malakar[1] · Soumya Deep Roy[2] · Soham Das[2] · Swaraj Sen[3] · Juan D. Velásquez[4,5] · Ram Sarkar[3]

Abstract

Disease prediction from diagnostic reports and pathological images using artificial intelligence (AI) and machine learning (ML) is one of the fastest emerging applications in recent days. Researchers are striving to achieve near-perfect results using advanced hardware technologies in amalgamation with AI and ML based approaches. As a result, a large number of AI and ML based methods are found in the literature. A systematic survey describing the state-of-the-art disease prediction methods, specifically chronic disease prediction algorithms, will provide a clear idea about the recent models developed in this field. This will also help the researchers to identify the research gaps present there. To this end, this paper looks over the approaches in the literature designed for predicting chronic diseases like Breast Cancer, Lung Cancer, Leukemia, Heart Disease, Diabetes, Chronic Kidney Disease and Liver Disease. The advantages and disadvantages of various techniques are thoroughly explained. This paper also presents a detailed performance comparison of different methods. Finally, it concludes the survey by highlighting some future research directions in this field that can be addressed through the forthcoming research attempts.

1 Introduction

In this digital era, many organizations have been established across the globe which provide continuous health monitoring facilities for humans. In the traditional method, patients visit the clinic and the health professionals advise them through their expertise in diagnosis. However, in this age-old way of medical diagnosis, patients face various difficulties owing to the increase in the number of health related problems as well as the population, especially in developing countries. This scenario sometimes leads to improper care of a patient, which can even prove fatal.

To this end, technology provides an alternative to the traditional system. Hence, it plays a significant role in healthcare systems by incorporating a large number of computer aided supporting systems and tools. This bonding has not only improved the quality of patient care but also reduced the cost of treatment by imparting efficient allocation of medical resources. The main components of technology-enabled healthcare systems are medical experts, hardware and software. However, designing an automatic system that can predict the disease from electronically available medical data is very challenging. The huge social impact of this research field motivates researchers from various domains like computer science, biology, medicine, statistics, and drug design. These researchers are continuously trying to come up with a near perfect system for better patient care.

In this context, it is worth mentioning that with the growing availability of digital records and data, the last two decades have observed an exhaustive adoption of data mining and machine learning (ML) techniques [89] in healthcare/

✉ Samir Malakar
  malakarsamir@gmail.com

  Soumya Deep Roy
  sdrjumme@gmail.com

  Swaraj Sen
  swaraj.sen2019@gmail.com

  Juan D. Velásquez
  jvelasqu@dii.uchile.cl

  Ram Sarkar
  raamsarkar@gmail.com

1 Department of Computer Science, Asutosh College, Kolkata, India

2 Department of Metallurgical and Material Engineering, Jadavpur University, Kolkata, India

3 Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

4 Departament of Industrial Engineering, University of Chile, Santiago, Chile

5 Instituto Sistemas Complejos de Ingeniería (ISCI), Santiago, Chile

patients care systems. Healthcare, one of the most crucial sectors of our society, is availing the facility of digitization of medical records and the emergence of Electronic Health Records (EHRs). The large-scale availability of EHRs has led to a surge in Computer aided Diagnosis and Detection (CADD) systems which, in general, employ various ML algorithms to accurately predict the presence of a particular disease in a subject. These CADD systems help in removing subjectivity in EHR and/or histopathology image analysis, thereby minimizing the prediction error. According to the third global survey[1] on electronic health (eHealth), conducted by the World Health Organization (WHO) in 2016, there has been a steady growth in the adoption of EHRs over the past 15 years and a 46.00% global increase in the past five years.

A disease is an abnormal condition that affects mostly a part of an organ and it is not caused by some external injury. In medical science, there are many categories of diseases like acute, infectious, heredity, and chronic. Chronic diseases generally persist for a longer period (3 months or more) in human organs. In general, such diseases cannot be prevented using vaccines or cured by medications. However, early detection of these diseases can save many human lives. Chronic diseases such as cancer, heart disease, and diabetes are the leading causes of death of human beings. With 784,821 deaths in India and an estimated 9.6 million deaths globally in 2018, cancer is considered as one of the most fatal diseases. Cancer generally involves abnormal cell growth with the potential to spread to other organs in the human body. The most common forms of cancer include Breast Cancer, Lung Cancer, Bronchial Cancer, Leukemia, Prostate Cancer, etc. It comes as no surprise that the earliest efforts in CADD [58] started with mammography for the detection of Breast Cancer, and these techniques were later applied to other forms of cancer as well.

Heart Disease, also called Cardiovascular Diseases (CVDs), is another group of diseases that causes a large number of deaths every year. Diseases that fall under the umbrella of CVDs include coronary artery disease, heart rhythm problems (also known as arrhythmias), and congenital heart diseases among others. According to Abdul-Aziz et al. [3], one in four deaths in India is due to CVD. According to a WHO report[2] more people (over 17.9 million each year) die from CVDs worldwide than from any other cause. Diabetes is one of the prime risk factors for CVDs. It is of two types: Diabetes insipidus or Type-I Diabetes in which the pancreas produces little to no insulin and Diabetes mellitus or Type-II Diabetes which affects the way the body processes insulin. Type-II Diabetes accounts for more than 90% deaths of all the Diabetes cases [155]. It mainly arises out of unhealthy lifestyle choices. Over 30 million people have been diagnosed with Diabetes in India and are attributed as the direct cause of 1.6 million deaths as of 2016[3].

Diabetes along with hypertension is responsible for Chronic Kidney Diseases (CKD). In CKD, the malfunctioning kidneys fail to filter waste from the blood leading to waste accumulation and may eventually lead to renal failure. A major impediment to CKD diagnosis is that the early stages of CKD show no symptoms. Around 100,000 patients are diagnosed with end stage kidney disease every year[4] in India. Unhealthy lifestyle choices also cause Liver Diseases. The liver is the largest organ in the body and Liver diseases mainly arise from excessive consumption of substances like alcohol, harmful gases, contaminated food, pickles and drugs [148]. With 259,749 deaths in 2017 in India and around 1.3 million deaths worldwide from cirrhosis alone, Liver Disease is one of the major health issues worldwide[5].

1.1 Motivation and Contributions

All the facts mentioned earlier demand the need for early detection of such chronic diseases, which could save a large number of human lives. As a result, several researchers around the globe engaged their time to find the solutions for detecting chronic diseases at their early stages. Also, the abundant data gathered from EHRs, medical diagnosis and medical imaging alongside the technological and technical improvement in ML and artificial intelligence (AI) have led to significant research in Bioinformatics, Biomedical imaging and CADD systems. There has been extensive research in the domain of disease prediction. Since 1997, such methods have been published in different refereed journals and conferences [163]. As a result, voluminous research articles are present in the literature. Accordingly, several review articles [5, 57, 59, 78, 91, 92, 92, 93, 103, 134, 177, 201] are also present in the literature on chronic disease prediction. We have summarized the types of methods these review articles considered in Table 1. From Table 1, we can observe that these articles were prepared by highlighting the problem domain narrowly i.e., these were prepared either by highlighting a specific chronic disease [90, 93, 93, 177] in most of the cases or in a few cases, a specific category of ML or AI-aided techniques [5, 91, 92, 103, 201]. Sometimes, the authors discussed specific ML or AI-aided techniques for

---

Table 1 Coverage of the present survey in comparison with the past surveys

Review article	Remarks
Fatima et al. [57]	1. Research works related to the prediction of heart, Diabetes, liver, dengue and hepatitis were included for discussion
	2. On average, 4 to 5 works for each disease were included for discussion
	3. No critical analysis was provided to highlight the future research directions for such diseases
Rigla et al. [177]	1. Research findings on Diabetes detection from diagnosis reports were considered
	2. Classical ML approaches like Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM) based methods were included
	3. Important factors like notable research findings, future research directions, critical analysis of the surveyed works were missing
Jain and Singh [103]	1. Importance was given to describing FS techniques rather than disease detection methods
	2. Most of the works used (more than 90%) for critical analysis were on Diabetes prediction
	3. Image based disease predictions were absent as all the cited works were on diagnosis report based repositories
	3. Disease specific discussion was missing
Ahmadi et al. [5]	1. Works considered for review are more fuzzy technique centric rather than disease-centric which means fuzzy techniques were prioritized rather than the disease detection
	2. Published papers till 2017 were considered for discussion
	3. Only 52 works on different (>10), i.e., on an average 3 to 4 works per disease, and diagnosis report based disease were used
Hosni et al. [91]	1. Seven research questions, mostly classifier ensemble oriented, were set, and the published works that address each of the research questions mentioned, were investigated. So this review is entirely focused on ensemble based methods
	2. Only the works on Lung Cancer disease detection from diagnosis reports were considered
	3. Ensemble methods that were considered followed classical ML algorithms
Hosni et al. [90]	1. Some of the notable research outcomes related to Breast Cancer detection from diagnosis reports were considered for analysis
	2. No description/critical analysis of the methods could be found. Rather categorical analyses of the works were made without individualization
	3. No description of used datasets and performance comparison and hence the beginners may find it difficult to get the complete idea over the problem domain
Fernández et al. [59]	1. Research initiatives for Diabetes disease detection were considered
	2. Disease detection techniques that employed classifier ensemble techniques were only included
	3. Better performing single classifier works were excluded based on the selection strategy of the research works
	4. No critical analysis of the works was provided
Hosni et al. [92]	1. Extended version of the work [91]
	2. Extension in terms of paper selection strategy and increase of one research question
	3. Statistics related to the questions were discussed. There was no focus on disease detection methods.
Gupta and Gupta [78]	1. Investigated the performance of Artificial NNs (ANNs), Restricted Boltzmann Machine, Deep Autoencoders, and Convolutional Neural Networks (CNNs) for screening Breast Cancer
	2. Described only 17 previously published research works
	3. Results on only the Surveillance, Epidemiology and End Results (SEER) dataset were investigated to suggest a suitable solution
Manhas et al. [134]	1. Reviewed image based cancer detection techniques
	2. Five types of cancer - Cervical, Oral, Breast, Brain and Skin were included in the review
	3. Diagnosis based cancer detection techniques were overlooked
Sharma and Rani [201]	1. A set of 12 research questions are identified for cancer research
	2. Gene expression data based research was highlighted most
	3. Research related diagnosis reports were also included
Houssein et al. [93]	1. Breast Cancer screening techniques were discussed
	2. Only research work that dealt with histopathological, mammogram images were included
	3. Diagnosis report based methods were also included

Table 1 (continued)

Review article	Remarks
Reshmi et al. [175]	1. Image based Breast Cancer detection techniques were included
	2. Only research works that dealt with histopathological images were included and thus minimized the research spectrum
	3. Other aspects of Breast Cancer detection were not discussed
Proposed, 2022	1. Several diseases that include Breast Cancer, Lung Cancer, Leukemia, Heart Disease, Diabetes, Liver Disease and CKD have been covered
	2. Research endeavors made over the last two decades involving both image based and diagnostic report based disease screening methods have been studied
	3. Various important ML strategies like missing value imputation, feature reduction, FS, classifier combination, fuzzy logic along with DL based strategies have been considered
	4. A comprehensive description of each work along with its shortcomings and possible scope for future improvement is provided. Also a comparative performance analysis of the methods for each disease has been provided
	5. The disease specific research trends are analyzed and accordingly some future research directions for each disease are suggested

a particular disease [78, 90, 177]. Hence, it can be safely concluded that although surveys published enlisting research efforts for specific diseases or specific ML techniques, such surveys fail to shed light on the prevailing trend of research across multiple diseases or ML approaches.

To this end, the present survey is a significant endeavor as it not only encapsulates the concerted research efforts in the specific diseases but also attempts to reflect on the current trends in research across chronic diseases like Breast Cancer, Leukemia, Lung Cancer, Heart Diseases, Diabetes, CKD and Liver, and AI based methods including missing value imputation, feature reduction, feature selection (FS), classifier combination, fuzzy logic, and ML and Deep Learning (DL) based approaches. This survey also chronicles the diagnostic report based and image based approaches for the said diseases. A comparative study of our survey with some recent review articles is reported in Table 1. Information in this table conveys that the present survey not only covers the similar methods described in state-of-the-art review articles but also includes a wider range of AI based applications in several disease prediction systems. It is to be noted that in the present article we only concentrate on automatic disease prediction systems (an integral part of technology-enabled health care systems) that use several ML and DL schemes for the prediction purposes.

1.2 Organization of the Article

The overall organization of this survey is shown in Table 2. Section 2 provides the detail working procedure of a generic disease prediction system and a generic diagnostic report based system and an image based system. Section 3 discusses the commonly applied evaluation metrics used for assessing the performance of a disease prediction system. Sections 4, 5, 6, 7 and 8 report the research endeavors in

Cancer, Heart Disease, Diabetes, Liver Disease and CKD respectively. Section 4 is subdivided into three subsections for describing Breast Cancer, Lung Cancer and Leukemia respectively, each of which further contains subsections for image and diagnostic report based disease prediction techniques. Whereas, Sect. 5 has separate sections for the single classifier based approach and the Ensemble based approach in Heart Disease detection. For an easy reference, the organization of disease specific discussions made in this article is shown in Fig. 1. Some important future research directions are discussed in Sect. 9. Finally, this survey is concluded in Sect. 10.

2 Generic Chronic Disease Prediction Method

A chronic disease prediction system accepts data from a new subject as input and generates some specific report about the status of the disease observed in the subject. In general, the status of the subject is labelled either positive or negative. Status is positive if the subject is affected by the disease for which the test is made. Otherwise, the subject is labelled negative. In some recent systems, the positive cases are further classified into more categories based on the stage of the disease. For example, in the Breast Cancer Histology (BACH) image dataset[6] Aresta et al. [17] labelled positive Breast Cancer stages into three stages: benign, in situ

---

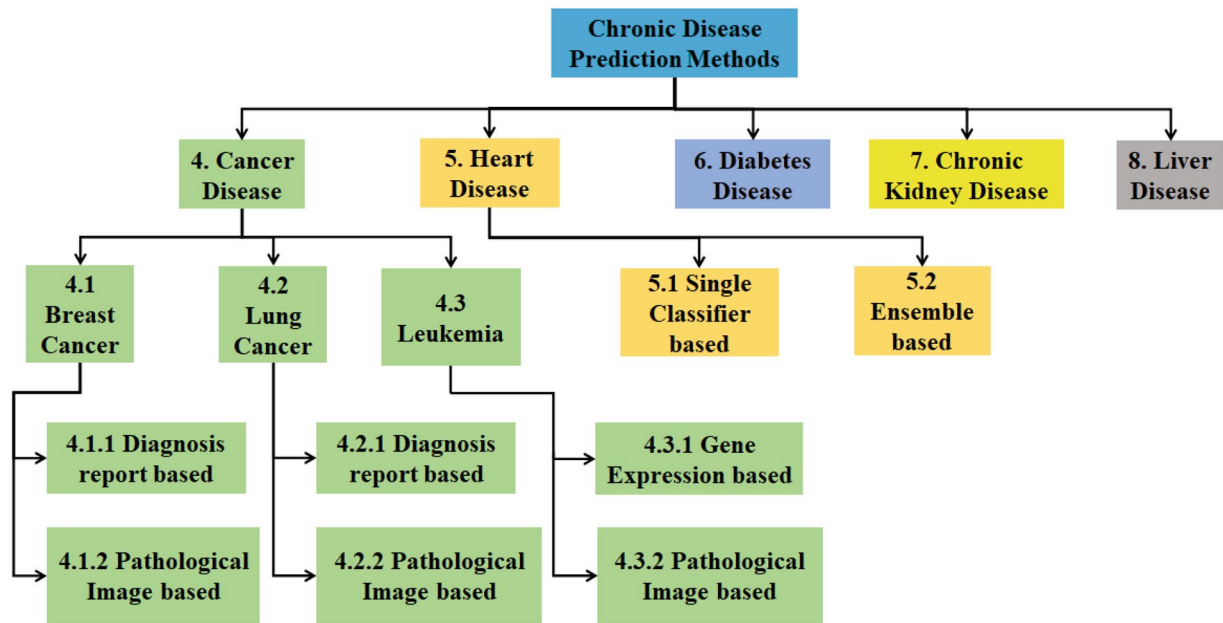[6] https://iciar2018-challenge.grand-challenge.org/Dataset/

Fig. 1 Organization of disease prediction methods discussed in this article. Numbers indicate sections and subsections where the particular topics are discussed

carcinoma and invasive carcinoma while in the Breast Cancer Histopathological Database (BreakHis)[7] positive Breast Cancer cases labelled into eight sub-categories.

In Fig. 2, we have provided a schematic diagram which shows how a generic disease prediction system works (here a chronic disease). The figure highlights the communication among subjects, doctors, pathological laboratories and AI assisted decision support systems. The clinical data (a set of diagnostic reports) for a subject may have three different forms: (i) input from the subject (e.g., height, weight, age, and sex, etc.), (ii) input from the doctors or experts (e.g., heart rate, body mass index, and disease specific observations, etc.), and (iii) input from a pathological laboratory (e.g., blood cell counts, pathological image(s) and X-ray image, etc.). It is noteworthy to mention that the diagnostic parameters vary from one disease to another. The collected clinical data are analysed by an AI-assisted disease prediction system and the system generates some forms of recommendations to be suggested by the doctor for the subject, which are here considered as the presence or absence of disease. The doctors analyse the recommendations and suggest treatment for the subject. In this survey, based on the nature of clinical data, we broadly divide the disease prediction systems into diagnosis report based systems and pathological image based systems.

2.1 Diagnosis Report based Method

In the early years of research related to computerized disease prediction systems, the patients' data (diagnostic reports) was mostly stored electronically in text form. As a result, a set of systems have been designed over time. In these systems, in general, a classifier is used to learn from these stored textual data. Most popular diagnosis report based datasets available electronically contains missing values, insignificant attributes and redundant information about the subject which do not help the classifiers to learn from these data sufficiently rather many times reduce their efficiency. Therefore, preprocessing techniques that include missing value imputation or data filtering (used as a substitute for missing value imputation) and data reduction have been employed by the researchers before feeding the data to a classification model. The preprocessed training samples are then fed to the selected classifier to generate a learning module that contains optimally tuned hyperparameters of the classifier. In the testing phase (or prediction phase), a diagnosis report is collected from an unknown subject and then preprocessed based on the available information from the training stage. Finally, the preprocessed data are fed to the saved learned model to decide whether the subject is carrying the disease or not. A generic model of diagnosis report based disease prediction system is shown in Fig. 3.

---

[7] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

Table 2 The overall organization of the article

Topic	Work Refs.

2.2 Pathological Image based Method

A diagnosis system that uses pathological images collected from the subjects to predict disease, in general, follows two different approaches: feature engineering based approach (see Fig. 4) and DL based approach (see Fig. 5). However, methods following any of these approaches commonly use some preprocessing techniques to obtain better features from the images. As shown in Figs. 4 and 5 after preprocessing, the general trend is to find out the Region of Interest (ROI) to pinpoint areas which can be used to extract the features. In the case of DL based methods, the high level features are extracted from ROI images through the use of a set of convolutional and pooling operations while in the case of the handcrafted features the domain knowledge contributes significantly while extracting features. In the case of feature
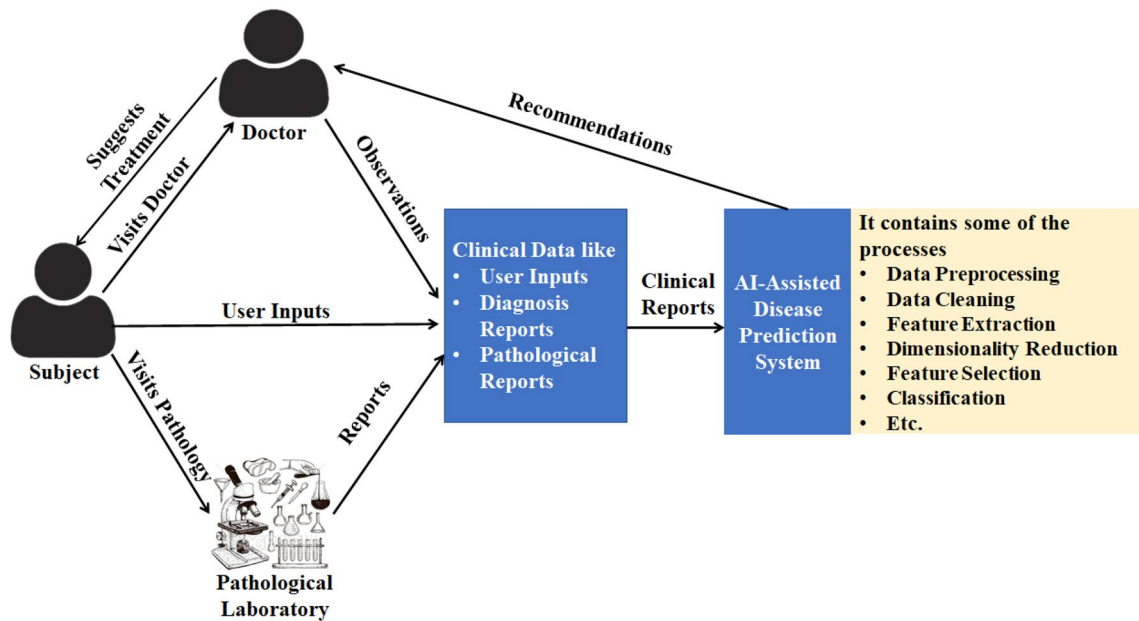
Fig. 2 Steps involved in a general decision support system

Fig. 3 A block diagram representing steps used, in general, for predicting a disease from a given set of diagnosis reports
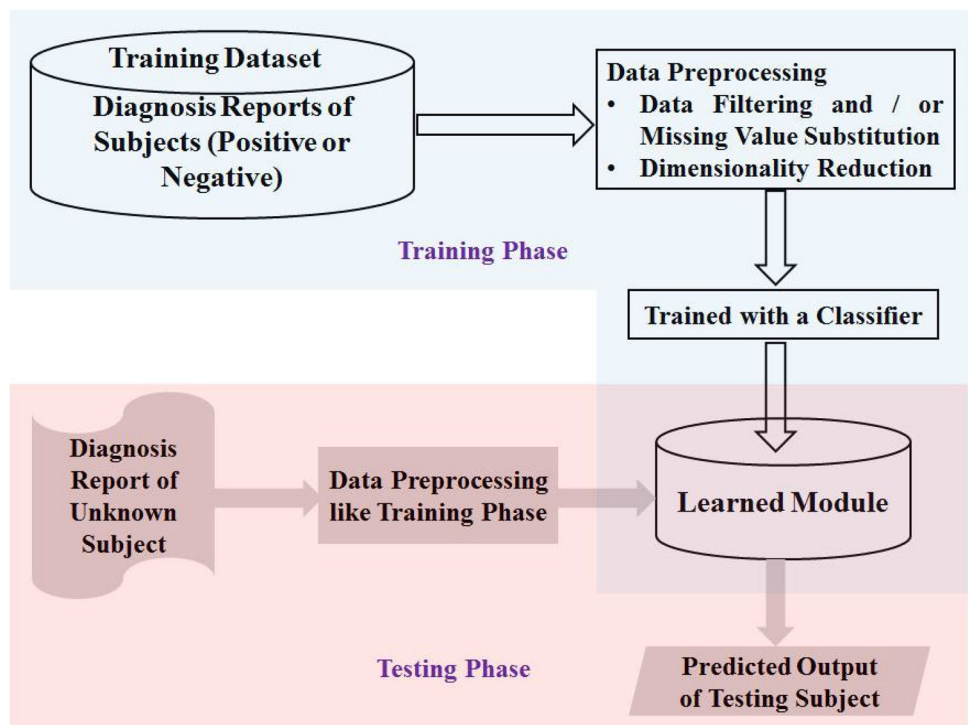
engineering based methods, many standard classifiers are used to predict the disease from unknown samples. Whereas in the case of DL based methods, it is generally performed using multi-layered NNs that automatically extract features from the input data.

3 Evaluation Metrics

The disease prediction systems described in this article used some classification models for disease diagnosis. Researchers used some standard metrics like accuracy (ACC), precision (P), recall (R), F1-score (F1), Specificity (S) and
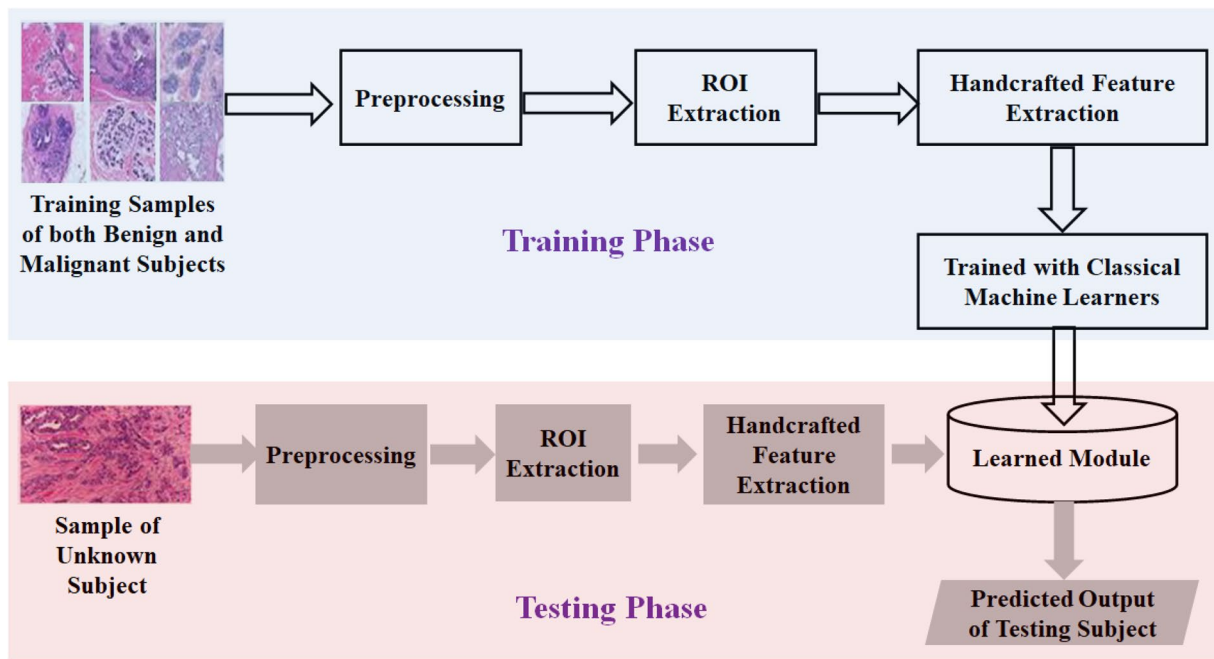
Fig. 4 A block diagram representing steps used, in general, for predicting a disease from a pathological image following handcrafted features extraction and shallow learner based classifiers
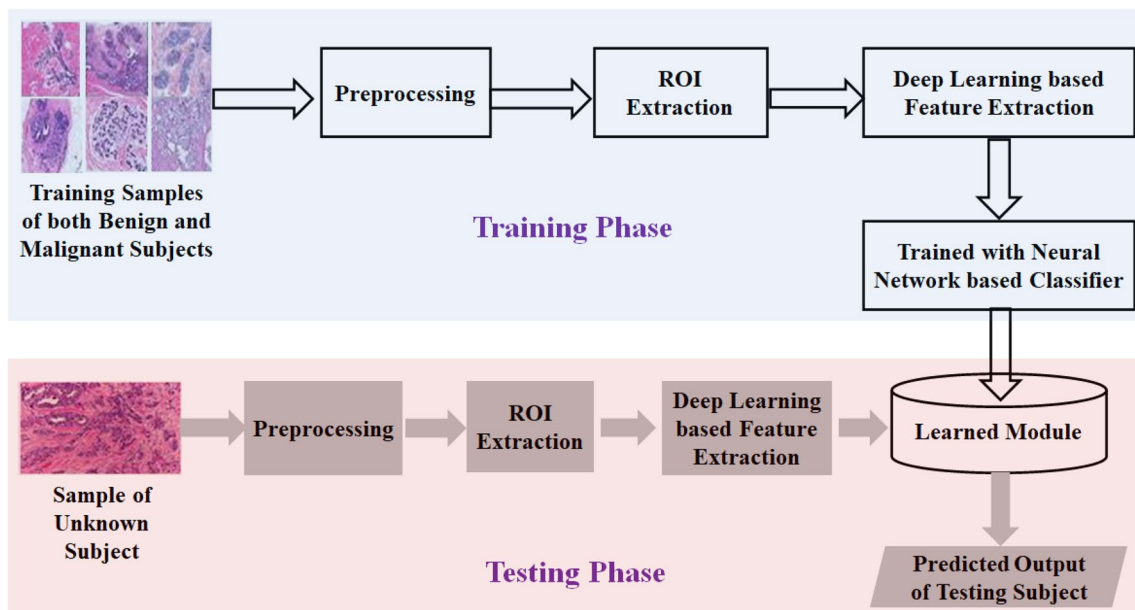
Fig. 5 A block diagram representing steps, in general, used for predicting a disease from pathological images using some DL based models

Area under Receiver Operator Characteristic (ROC) Curve (ROC-AUC score or simply AUC score) of a classification model while citing the performance of their model. These evaluation metrics can easily be measured using the confusion matrix or error matrix which describes the complete performance of a model while predicting the result. It is a $N \times N$ (N is the number of classes) dimensional matrix. It helps to summarize and visualize the performance of a predictive model. It consists of four metrics: True Positive (TP),

True Negative (TN), False Positive (FP), and False Negative (FN), which are defined below.

- **TP** indicates the number of positive samples that are predicted correctly by the classification model
- **TN** indicates the number of negative samples that are predicted as negative samples by the classification model.
- **FP** indicates the number of negative samples that are predicted as positive samples by the classification model
- **FN** indicates the number of positive samples that are predicted as negative samples by the classification model

3.1 Accuracy

One of the most commonly used metrics for reporting the performance of a disease prediction method is accuracy. It indicates how often a disease prediction model correctly predicts a sample. It is calculated as the fraction of the number of correctly classified samples (i.e., $TP + TN$) and the total number of test samples (i.e., $TP + TN + FP + FN$) and it is defined by Eq. 1.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

If the test set suffers from the imbalanced class problem [32, 131] which is very common for medical datasets as normal cases overwhelm disease cases then this metric may mislead the overall model performance. Thus, the use of other metrics like P, R, F1, S, and AUC scores become necessary and thus used widely in the past works.

3.2 Precision

Precision, also known as Positive Predictive Value (PPV), represents the classifier's capability to predict positive cases as the positive. It is calculated using Eq. 2.

$$P = \frac{TP}{TP + FP} \tag{2}$$

3.3 Recall

Recall, alternatively known as sensitivity, is the fraction of positive samples that are classified by the model correctly with respect to the total number of actual positive cases present in the test dataset. Recall is calculated using Eq. 3.

$$R = \frac{TP}{TP + FN} \tag{3}$$

3.4 F1-Score (F1)

F1-score represents the trade off between recall and precision scores, and the harmonic mean of precision and recall is considered as the F1-score value. When the FP and FP are equally important then the F1-score metric is very useful in order to know about the model's prediction capability. It is calculated using Eq. 4.

$$F1 - Score = 2 * \frac{P * R}{P + R} \tag{4}$$

3.5 Specificity

Specificity, also known as TN rate, is the proportion of negative samples that are correctly classified i.e., it is the fraction of TN and $TN + FP$. It is calculated using Eq. 5.

$$S = \frac{TN}{TN + FP} \tag{5}$$

3.6 ROC-AUC Score (AUC)

ROC curve is an evaluation metric for classification problems and disease diagnostic tasks, where the curve represents the probability curve that plots the TP rate against the FP rate at various threshold values. It is used to show the trade-off between sensitivity and specificity in a graphical way. In other words, the AUC score describes how an adjustable threshold causes changes in two types of errors: false positives and false negatives. AUC score works as a quantitative measure of the ROC curve based evaluation metric. It tells us how much a model is capable of distinguishing between classes. A higher value of AUC means the model is better at class prediction. The readers are suggested to read the article by Carrington et al. [32] for a more insightful explanation of the same.

4 Cancer Prediction Methods

In the current section, we have mostly discussed recent DL and ML aided cancer detection methods. Here, three types of cancer viz., Breast Cancer, Lung Cancer and Leukemia prediction systems have been discussed.

Table 3 Performance comparison of recent prediction models on WBC datasets

Work Ref.	Technique Used	Performance metrics				
		ACC	P	R	F1	S
Zheng et al. [234]	K-means aided feature transform + SVM	0.94	–	0.96	–	0.92
Kamel et al. [109]	GWO FS + SVM	1.00	–	1.00	–	1.00
Mafarja et al. [129]	Hybridized WOA and SA	0.97	–	–	–	–
Ghosh et al. [66]	SMO-X FS + k-NN	1.00	–	–	–	–
Huang et al. [98]	GA-RBF SVM (small scale data)	0.98	–	–	0.99	–
	RBF-SVM (large scale data)	0.99	–	–	0.99	–
Kumari and Singh [122]	k-NN	0.99	–	–	–	–
Kumar et al. [121]	Lazy k-star/ Lazy IBK.	0.99	0.99	0.99	0.99	–
Prakash and Rajkumar [165]	HLFDA-T2FNN	0.98	–	–	–	–
Preetha and Jinny [166]	PCA-LDA + ANNFIA	0.97	–	–	–	–
Sandhiya and Palani [186]	ICRF-LCFS	0.94	–	–	–	–
Chatterjee et al. [36]	SSD-LAHC + k-NN	0.99	–	–	–	–
Ahmed et al. [6]	RTHS + k-NN	0.99	–	–	–	–
Guha et al. [72]	CPBGSA +MLP	0.99	–	–	–	–
Guha et al. [73]	ECWSA + k-NN	0.95	–	–	–	–
Ghosh et al. [67]	MRFO + k-NN	1.00	–	–	–	–

4.1 Breast Cancer Prediction Methods

We have elucidated some of the existing Breast Cancer disease detection techniques present in the literature. We have come across some quality research works related to both diagnosis reports and image based methods for Breast Cancer detection which are discussed hereafter.

4.1.1 Diagnosis Report based Methods

In literature, a significant number of research articles used the Wisconsin Breast Cancer (WBC)[8] dataset in Breast Cancer detection research. It contains a diagnosis report constituting 9 attributes of 699 subjects (241 positive subjects and 458 negative subjects). The values of all these attributes are scaled to a range of 1 to 10 depending on their proportion. A class attribute (value is 2 for benign tumor and 4 for malignant tumor) is also present in the dataset. By thorough analysis of the state-of-the-art methods, we have observed that there are many works on this dataset using different ML based approaches. However, in most of the cases, methods involving SVM have outperformed the rest. Table 3 lists the performance of various research efforts on the WBC dataset. Some of the important methods are discussed here.

The FS based techniques are mostly employed in the last 10 years for Breast Cancer detection while considering the WBC dataset. In 2014, Zheng et al. [234] designed a feature transform method using the k-means algorithm where they first generated n (< number of attributes) possible cluster centres using the k-means algorithm on the training dataset and then the distances of test samples from these n cluster centres were considered as transformed features. Later to perform the disease detection, the transformed feature vectors were classified using the SVM classifier. The authors have proved experimentally that it outperformed the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) based FS techniques while using SVM as a classifier. Kamel et al. [109] showed the importance of a good FS technique for this research problem. In this work, outliers were eliminated utilizing the outer line method. They utilized Grey Wolf Optimization (GWO) [141] for FS. This FS technique yielded an extraordinary result using the SVM classifier. In another work, Mafarja et al. [129] proposed a hybrid FS approach, where the authors utilized the Whale Optimization Algorithm (WOA) [140] and Simulated Annealing (SA) [116]. The authors obtained very promising results for FS where only 4 attributes from 9 were selected. In another work, Fruit Fly Optimization Algorithm (FOA) [156] was used by Shen et al. [204] for selecting diagnostic attributes. FOA iteratively obtains the optimal features. Moorthy and Gandhi [149] proposed another hybrid FS technique using Analysis of Variance (ANOVA) and WOA based FS techniques. With this model, the accuracy of the dataset with the SVM classifier increased by 2.10% while compared to SVM based classification without FS.

Ghosh et al. [66] introduced a swarm based FS technique named social mimic optimizer using an X-shaped transfer function (called SMO-X). With SMO-X, the authors gained the maximum accuracy with the k-nearest neighbors (k-NN) classifier with the help of 33.33% of actual features.

In another work, Prakash and Rajkumar [165] used a hybrid FS technique termed Hybrid Local Fisher Discrimination Analysis (HLFDA) [61]. After the FS, the dataset that has a near optimal feature subset was fed into Type 2 Fuzzy NN (T2FNN) for classification. Preetha and Jinny [166] combined two powerful feature transform strategies: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). They put the dataset individually in PCA and LDA first and then combined the resultant transformed features to obtain the final feature set, which was used for disease detection. Performances of three classifiers viz., Adaptive Neuro-Fuzzy Inference Systems (ANFIS), SVM and Multilayer Perceptron (MLP) were investigated and out of which ANFISs outperformed others. Recently, Dey et al. [49] proposed a hybrid FS technique called Learning Automata based Grasshopper Optimization Algorithm (LAGOA) where the premature convergence of Grasshopper Optimization Algorithm (GOA) was taken care of. The Learning Automata (LA) was used to adjust GOA parameters adaptively while a two-phase mutation technique was utilized to increase the exploitation capability of GOA.

One can also find a number of hybrid FS methods [6, 36, 67, 72, 73] where the researchers employed their method on WBC dataset. For example, Chatterjee et al. [36] designed a hybrid FS method which improved the local search capability of the Social Ski Driver (SSD) algorithm with the help of the Late Acceptance Hill Climbing (LAHC) method. The method is named as SSD-LAHC algorithm. In another work, Ahmed et al. [6] proposed Ring Theory based Harmony Search (RTHS) algorithm which coupled well-known Harmony Search (HS) algorithm and Ring Theory based Evolutionary Algorithm (RTEA). Guha et al. [72] designed a Clustering based Population in Binary Gravitational Search Algorithm (CPBGSA). In this work, the authors decided initial population using a clustering method to overcome the premature convergence of the Gravitational Search Algorithm. In another work, Guha et al. [73] proposed Embedded Chaotic Whale Survival Algorithm (ECWSA) using Whale Optimization Algorithm (WOA) aiming at better classification accuracy. They used a filter method to refine the selected features by WSA. Ghosh et al. in [67] compared the performance of Manta Ray Foraging Optimization (MRFO) with varying transfer functions. In this work, the performances of four variants of S-shaped and V-shaped transfer functions were studied.

Apart from the FS based techniques, we also find the use of other classical ML approaches. For example, Huang et al. [98] performed a detailed comparative study between SVM with the varying kernel (i.e., Liner, Polynomial and Radial Basis Function (RBF)) and ensemble methods (bagging and boosting) having SVM as classifier. They also experimented with GA to establish the effect of FS on classification in the WBC dataset. Through extensive experiments, they established that in the case of small datasets, boosted SVM with polynomial kernel and GA performed the best while for large datasets, SVM with RBF kernel and SVM with polynomial kernel performed the best. Missing values in data can put a detrimental effect on the classifier. To ameliorate this, Choudhury and Pal [43] introduced a novel method of missing value imputation using autoencoder NN. In this method, first the autoencoder NN was trained on multiple datasets like WDBC without any missing values for the attributes present there and then the learned autoencoder was used to predict missing values for other datasets that contain some missing values. As an initial guess of the missing value, the nearest neighbor rule was used and then refined by minimizing the reconstruction error. This was based on the hypothesis that a good choice for a missing value would be the one that can reconstruct itself from the autoencoder. Recently, Shaw et al. [202, 203] used methods to handle imbalance class problems and subsequently improved the cancer prediction performance. In the work [203], the authors proposed an ensemble approach to handle the class problem while in [202] they solved the problem using an evolutionary algorithm. In [202], the ring theory-based algorithm was hybridized with the PSO algorithm to select the near-optimal majority class samples from the training set. Such an initiative improved the final Breast Cancer detection performance of the classifiers. Apart from these major works, many authors like Kumari and Singh [122] and Kumar et al. [121] have given prominence to comparative studies. Kumari and Singh [122] applied correlation based FS and established the superiority of k-NN to SVM and LR classifiers while Kumar et al. [121] compared 12 classifiers, namely Adaptive Boosting (AdaBoost) algorithm, decision table, J-Rip, a DT classifier (J48), k-NN, Lazy K-star, LR, Multiclass Classifier, MLP, NB, RF and Random Tree (RT). The authors recommended the use of four classifiers, namely RF, RT, Lazy k-star, and k-NN that yielded the best accuracy during their experiments.

Performances of some state-of-the-art works on the WBC dataset have been recorded in Table 3. These results depict that the better performing methods used FS techniques before classifier training. Also, the methods like Shaw et al. [202], Ghosh et al. [67], and Dey et al. [49] provide state-of-the-art results on UCI Breast Cancer Wisconsin (Diagnostic) Dataset[9]. Hence, we can say that the use of the FS method will be considered a standard norm before using a classifier for diagnosis report based disease prediction systems.

---

[9] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

(a) Ductal carcinoma

(b) Lobular carcinoma

(c) Adenosis

(d) Fibroadenoma

(e) Phyllodes tumor

(f) Tubular adenoma

(g) Mucinous carcinoma

(h) Papillary carcinoma
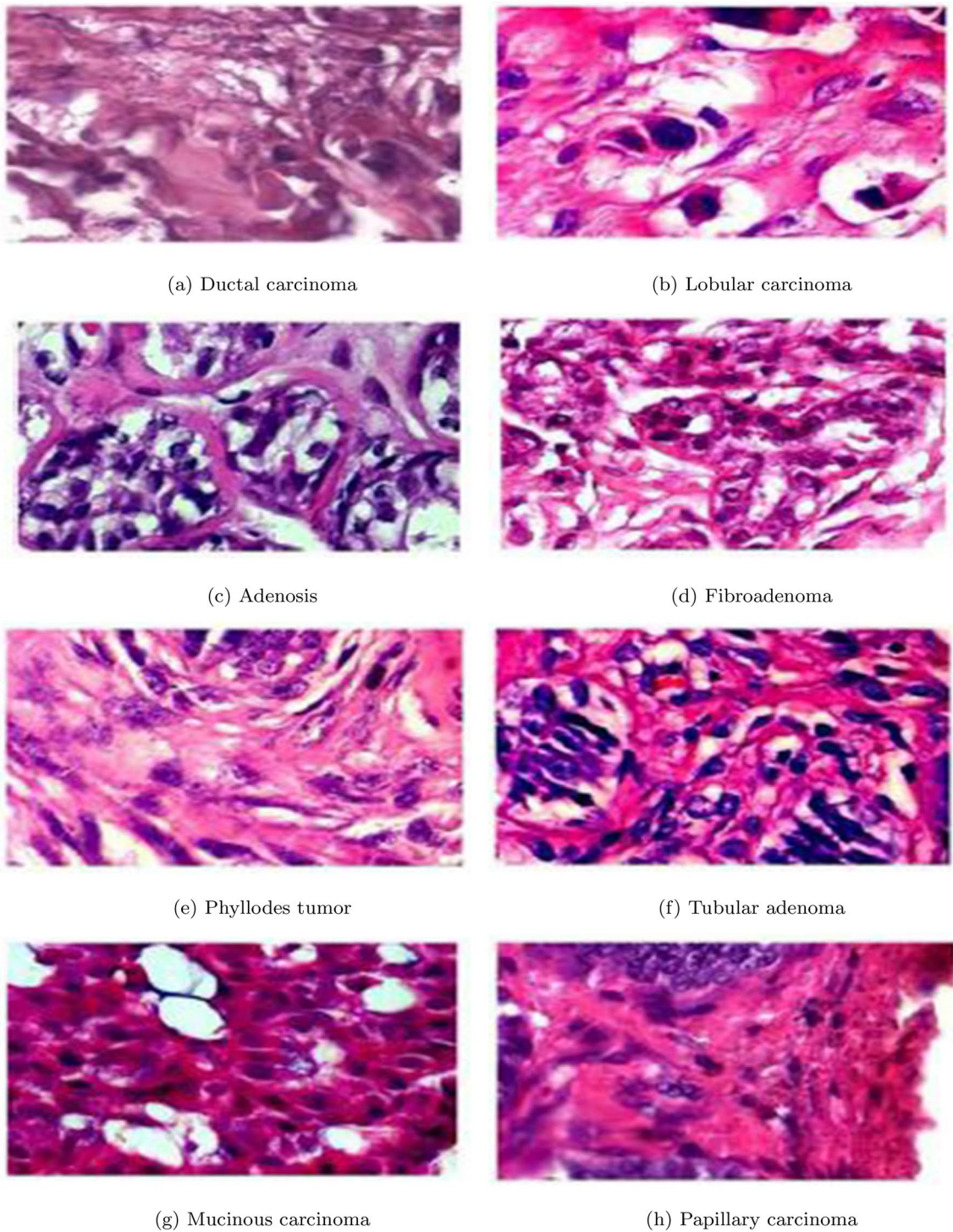
Fig. 6 Sample images from BreakHIS dataset : Eight classes of Breast Cancer histopathological images

4.1.2 Pathological Image Based Methods

It has already been mentioned that Breast Cancer is a deadly disease and early detection of the same is very crucial. Various diagnostic methods, including physical examination, ultrasound, magnetic resonance imaging, mammography, and biopsy have been developed and employed in the past. Among these, the histopathological image based diagnosis resulting from needle biopsy has turned out to be the gold standard in diagnosing Breast Cancer. However, DL and ML

(a) 40x

(b) 100x

(c) 200x

(d) 400x
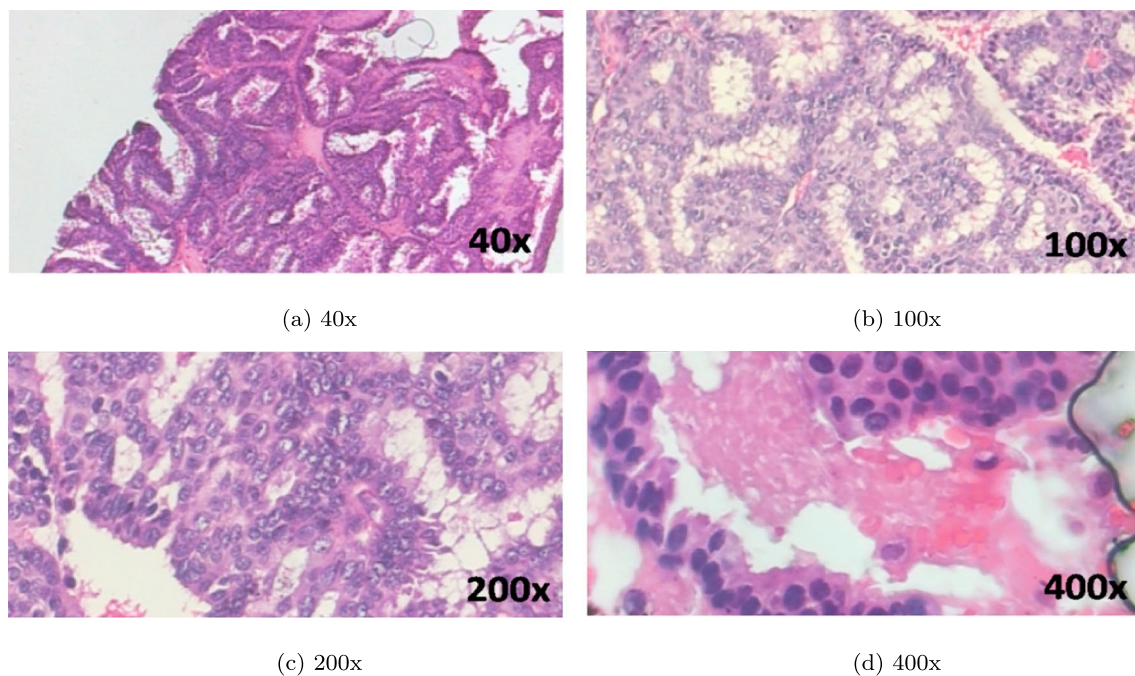
Fig. 7 Sample images from BreakHIS dataset: Papillary Carcinoma (malignant) tumor seen at different magnification levels

based CADDs are steadily finding their way into diagnostic studies. Deep NN (DNN) is approaching or even surpassing the accuracy and precision of histopathological diagnosis conducted by pathologists when identifying important features in diagnostic imaging studies.

The DL and ML aided research works on Breast Cancer prognosis considered here have proposed histopathological image based classification while using the BreakHis and BACH datasets mostly for conducting experiments. The BreakHis dataset is made up of 9,109 histopathological images (2,480 benign and 5,429 malignant images) of breast tumor tissue. Images have been amassed from 82 patients using four different magnifying factors (40X, 100X, 200X, and 400X). Each image has been stored in PNG format with a resolution of (700 × 400) pixels. Some samples from this dataset are shown in Figs. 6 and 7. The BACH dataset contains microscopy images that are made available through the grand challenge under International Conference on Image Analysis and Recognition (ICIAR 2018). This dataset consists of 400 Hematoxylin and Eosin (H&E) stained breast images of dimension (2048 × 1536) labelled with one of the four classes, namely normal, benign, in situ carcinoma and invasive carcinoma. Images have been annotated by medical experts. Some samples from the BACH dataset are shown in Fig. 8.

In the last few years, CNNs have undergone a substantial amount of tweaking and evolution. This has left us with a plethora of CNN architectures like VGG-16 [205], Inception-v1 [214], Inception-v3 [215], ResNet-50 [85], Xception

[42] and DenseNet [96], all of which were extensively used in image based Breast Cancer prediction. Shallu and Mehra [199] presented a comparative study employing three pre-trained networks, VGG-16, VGG-19 and ResNet-50 for fine-tuning and full training. During the preprocessing of BreakHis images, data augmentation was carried out by rotation of images about their centre. The effect of image scaling and the influence of training data size on overall system performance was emphasized using fully trained (where weights were randomly assigned) and fine-tuned CNNs. However, overfitting caused by the enormous capacity of the network was not rectified using the freezing layers due to space constraints. The fine-tuned pre-trained VGG-16 model based feature extraction along with the LR classifier outperformed the rest of the pre-trained networks. In another work, Yan et al. [228] proposed a hybrid model involving CNN and recurrent NN (RNN) models. They released a larger and more diverse dataset[10] consisting of 3771 Breast Cancer high resolution and annotated H&E stained histopathological images. Each image was labelled as normal, benign, in situ carcinoma, or invasive carcinoma. Instead of using the entire dataset, the dataset they used had the same size as the BACH dataset. The authors prepared a study and comparative analysis report on different combinations of patch-wise and image-wise DL based methods. In the pre-processing stage, patches were extracted from each image

---

[10] http://ear.ict.ac.cn/?page_id=1616

followed by data augmentation. A fine-tuned Inception-V3 was employed to extract features from each patch which were then fed into a bidirectional long short-term memory (BLSTM) to coalesce the features.

Coming to ResNets, Jannesari et al. [105] investigated the performances of ResNet extensions and all versions of the Inception model (i.e., v1-v4). Data augmentation methods like flipping, rotating, cropping and random resizing were applied to BreakHis dataset images. Saturation, brightness and contrast were applied to incorporate color distortions. Fine-tuned ResNet-152 competently classified malignant and benign cancer types with the highest accuracy. Unlike many studies where a single magnification level was used, this work used all the magnification levels i.e., 40X, 100X, 200X and 400X of BreakHis data. Rakhlin et al. [172] extracted deep features using three pre-trained CNN networks, namely ResNet-50, Inception-v3 and VGG-16 from strain normalized images. To encounter the overfitting problem the authors utilized an unsupervised dimensionality reduction mechanism along with data augmentation like image scaling and cropping. Recently, Gupta and Chawla [77] used pre-trained CNN models, namely, VGG-16, VGG-19, Xception, and ResNet-50 to extract features from histopathological images. CNNs were trained separately on different magnification factors (40x, 100x, 200x and 400x). The BreakHis dataset was used for performing experiments. The top layers i.e., the FC layers of the CNN model were removed and replaced by traditional ML methods i.e., SVM and LR. Features from pre-trained ResNet-50 with LR classifier outperformed other models on 40x and a 100x magnification factor. In another work, Dey et al. [52] used a pre-trained DenseNet-121 model for feature extraction purposes to detect Breast Cancer from thermal images. However, instead of passing the original image to the CNN model, they preprocessed the image using edge detection methods. The authors stacked edge images generated by Prewitt and Roberts edge detection technique with the original gray-level image to generate a three channel image. On the other hand, Sánchez-Cauce et al. [185] proposed a multi-input CNN model for detection of breast cancer from multi-modal data consist of patients' information along with thermal images.

The above mentioned methods utilized the existing CNN architectures to obtain better results. However, some instances are found where researchers designed their own CNN architecture instead of using the popular CNN architectures as mentioned earlier. For example, Han et al. [79] designed Class Structure-based Deep CNN (CSDCNN) for the classification of Breast Cancer. Training instances were augmented to get rid of the imbalanced class problem. In their experiments, they made use of two different training strategies. In the first case, they trained the CSDCNN from scratch on the BreakHis dataset while the second case was based on transfer learning that initially pre-trains CSDCNN

on ImageNet-37, followed by fine-tuning on BreakHis dataset. Sudharshan et al. [212] shed light on Multiple Instance Learning (MIL) and Single Instance Classification (SIC) problems while dealing with the Breast Cancer detection problem. This work divulged that MIL allows obtaining comparable or better results than SIC without labeling all the images. In this work, MIL methodologies like Axis-Parallel Hyper Rectangle (APR), diverse density, citation-k-NN, and SVMs with linear, polynomial and RBF kernels, non-parametric MIL and Multiple Instance Learning based CNN (MILCNN) were applied to the dataset. For each methodology, a grid search was used to tune their respective hyperparameters. After successive evaluations, it transpired that the non-parametric MIL and MILCNN-APR performed adroitly and yielded the best results while APR and citation-k-NN did not perform up to the mark.

It is well-known that the performance of weak classifiers can be improved by using classifier ensemble techniques. Thus, many researchers in the literature have designed classifier combination techniques for image based Breast Cancer detection. For example, Yang et al. [230] proposed the Ensemble of Multi Scale CNN (EMS-NN) model. To form the ensemble, they utilized pre-trained models of DenseNet-161, ResNet-152 and ResNet-101. The proposed methodology followed three stages, the first being, Multiscale image patch extraction followed by Training multiple DCNNs and Model selection and combination. Overfitting was attenuated by applying various data augmentation techniques that were used to inflate the size and color diversity. However, Chennamsetty et al. [40] used two pre-trained CNN models - DenseNet-161 and ResNet-101 that were employed on differently preprocessed histology images. Brancati et al. [30] designed their ensemble model with different versions of the ResNet CNN architectures. The notable contribution of the work is that the authors reduced the problem complexity using a down-sampling technique where the image size was reduced by a factor of k. It also used the central patch of size $m * m$ as input to the model which reduced the training complexity of the model further. Recently, Bhowal et al. [28] proposed a classifier combination method that used Choquet fuzzy integral as the aggregator function. The aggregator was used to combine the confidence scores returned by the CNN based classifiers. The most notable aspect of the work is that the authors utilized coalition game and information theory to estimate the fuzzy measures used in Choquet integral with the help of validation accuracy. In another work, Chouhan et al. [44] proposed emotional learning inspired feature fusion strategy to classify a mammogram image into normal and abnormal classes. The authors used three static features: taxonomic indices, statistical measures and LBP along with CNN based features. Chattopadhyay et al. [38] designed a DL based method called Dense Residual Dual-shuffle Attention

Table 4 Performance comparison of recent prediction models on various Breast Cancer datasets. Classifiers with '*' represent the classifiers with the highest ACC scores in the respective paper

Work Ref.	Classifier	Performance metrics					
		ACC	P	R	F1	S	AUC
BreakHis Dataset							
Han et al. [79]	CSDCNN	0.93	–	–	–	–	–
Sudharshan et al. [212]	MILCNN*	0.92	–	–	–	–	–
Shallu and Mehra [199]	VGG 16 + LR	0.92	0.93	0.93	0.93	–	0.95
Jannesari et al. [105]	ResNet V1 152	0.98	0.99	–	–	–	0.98
Gupta and Chawla [77]	ResNet50+LR	0.93	–	–	–	–	–
Chattopadhyay et al. [38]	DRDA-Net	0.98	–	–	–	–	–
BACH Dataset							
Rakhlin et al. [172]	LightGBM + CNN	0.87	–	–	–	–	–
Yang et al. [230]	EMS-Net	0.91	–	–	–	–	–
Roy et al. [178]	Self-designed (OPOD)	0.77	0.77	–	0.77	0.77	–
Roy et al. [178]	Self-designed (APOD)	0.90	0.92	–	0.90	0.90	–
Sanyal et al. [189]	Hybrid Ensemble (OPOD)	0.87	0.86	0.87	0.86	0.99	–
Sanyal et al. [189]	Hybrid Ensemble (APOD)	0.95	0.95	0.95	0.95	0.98	–
Bhowal et al. [28]	Choquet fuzzy integral and coalition game based classifier ensemble	0.95	–	–	–	–	–
Mics. Dataset							
Yan et al. [228]	Inception-V3*	0.91	–	0.87	–	–	0.89
Dey et al. [52]	DenseNet-121	0.99	0.99	0.98	0.98	–	–
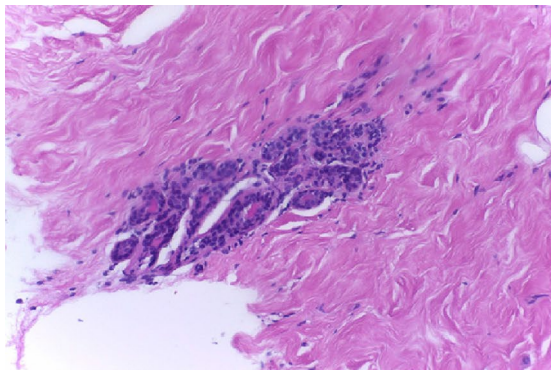
(a) Normal
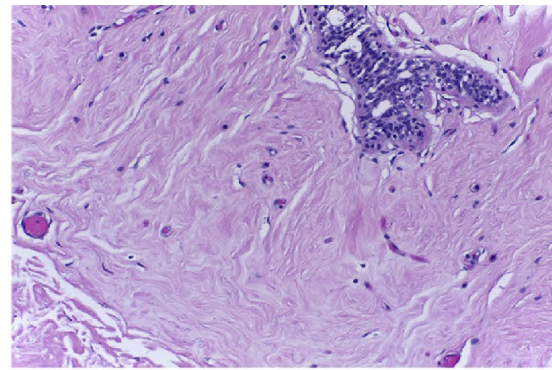
(b) Benign
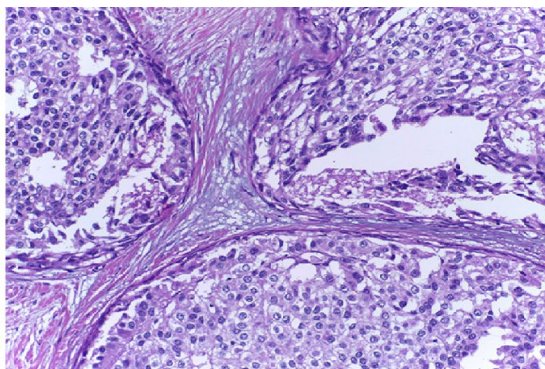
(c) In situ
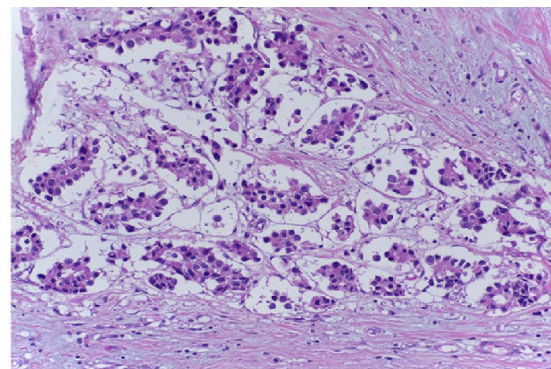
(d) Invasive

Fig. 8 Examples of different forms of microscopic biopsy images taken from the BACH dataset

Network (DRDA-Net) for detection of Breast Cancer detection from histopathological image.

All the methods described above used the entire image to detect the presence of Breast Cancer using the pathological image i.e., all these methods passed the entire image to the detection system. However, multi-view analysis of the pathological images might help in improving the final result. With this objective, some authors followed patch-based approaches for the said task. For example, Roy et al. [178] designed a CNN aided patch-based classification model for classifying histology breast images. Patches carrying distinguishing information were extracted from the original images to perform classification, which followed two different evaluation strategies, namely One Patch in One Decision (OPOD) and All Patches in One Decision (APOD). In another work, Sanyal et al. [189] proposed a similar detection like Roy et al. [178]. However, in this work, the authors utilized a novel hybrid ensemble approach for classification purposes while Roy et al. [178] used a single CNN architecture. This ensemble model used the confidence score returned by base CNN models and the confidence scores returned by the Extreme gradient boosting trees (XGB) classifier with the help of different features to make the final decision at the patch level. Finally, the patch level results are again ensembled to take the image level decision.

All the above discussed works mostly used deep features that might contain some redundant/irrelevant features. Therefore, the use of some FS methods might help in improving the final output. Based on this, Chatterjee et al. [37] proposed a deep feature selection technique. In this method, the authors improved the Dragonfly Algorithm (DA) with the help of the Grunwald-Letnikov method to perform FS. By doing this the authors improved the model's performance with fewer features. In computer vision, it is also observed that fusing handcrafted features with deep features improves the overall performance. Relying on this observation, Sethy and Behera [195] first concatenated LBP features with deep features extracted using the VGG19 CNN model and then classified them using the K-NN classifier. The performances of some important state-of-the-art Breast Cancer detection methods on different datasets are summarized in Table 4. The performance of state-of-the-art methods on different datasets is satisfactory but considering its deadliness characteristics it remains an open research problem. After analysing the performance trends of the methods studied here, we can safely comment that patch based methods assisted with DL performed better on BACH dataset. So, in the future, this approach could be studied on other datasets even for designing a better Breast Cancer detection technique from pathological images.

4.2 Lung Cancer Prediction Methods

Like Breast Cancer, Lung Cancer detection methods are also divided into two categories in this work: diagnosis report based and pathological image based methods. Some research attempts produced satisfactory results with ML algorithms like SVM and AdaBoost. However, in most of the works, clear domination of DL based techniques is found for image based as well as diagnosis report based methods.

4.2.1 Diagnosis Report based Methods

The dataset that is very commonly used in diagnosis report based Lung Cancer detection is from UCI-Irvine repository[11]. It contains 32 instances of data where the number of attributes is 57 (1 class attribute and 56 predictive attributes). It is an age-old dataset and contains a very limited number of samples with a large number of diagnosis attributes makes it a challenging dataset. However, some works used Lung Cancer Data[12] (LCD), survey Lung Cancer Data[13] (SLCD) of data from world repository, Thoracic Surgery Dataset (TSD) of UCI repository [23], Michigan Lung Cancer Dataset[14] (MLCD) and SEER[15] dataset. LCD dataset contains 1000 instances with 24 attributes having three different class labels: '0' is for a healthy person, '1' for a person with a benign tumor and '2' for a person with a malignant tumor. SLCD contains 309 instances but with 16 attributes. Out of these 16 attributes, 14 attributes have values 1 and 2 to represent NO and YES respectively while for the gender M (for male) and F (for female) are used. The values of attribute age are normalized by the min-max normalization method. The TSD contains 17 attributes (14 having nominal values and 3 having continuous values) and 470 instances. In this dataset class label, the attribute is represented by RiskYr which is a binary attribute. Risk1Yr is true if the person is dead and false if the person is alive. MLCD contains 96 instances (86 instances are cancerous and 10 are healthy) and 7130 attributes. SEER dataset is a gene expression based dataset which was released in April 2016. It contains 149 attributes and 643,924 instances. The dataset named lungdata lists the performance of various research efforts on various datasets discussed here. The non-availability of the diversified dataset for Lung Cancer disease detection for a long period might lessen the quality of research works as compared to other cases.

Table 5 Performance comparison of recent prediction models on various Lung Cancer datasets

Work Ref.	Classifier	Performance Metrics					
		ACC	P	R	F1	Error	AUC
UCI -TSD Dataset							
Danjuma [47]	MLP	0.82	0.82	0.82	0.82	–	0.84
MLCD Dataset							
Murty et al. [151]	RBF-NN	–	–	–	–	0.19	–
UCI-Irvine Dataset							
Patra [161]	RBF-NN	0.81	0.81	0.81	0.81	–	0.75
Dey et al. [49]	LAGOA+RF	0.86	–	–	–	–	–
LCD Dataset							
Radhika et al. [169]	SVM	0.99	–	–	–	–	–
Salaken et al. [183]	Deep-ANN	0.80	–	–	–	–	–
Maleki et al. [132]	Deep-ANN	1.00	–	–	–	–	–
SEER Dataset							
Ali and Reza [180]	RIPPER with AdaBoost	0.89	–	–	–	–	0.95
Doppalapudi et al. [53]	ANN	0.71	0.71	0.71	0.71	–	0.87
SLCD Dataset							
Nasser and Abu-Naser [153]	ANN	0.97	–	–	–	–	–

Most of the works in the literature performed experimental comparative studies consisting of different shallow classifiers. For example, Danjuma [47] compared the performance of three different classifiers: MLP, J48 and Naive Bayes (NB) on TSD from the UCI repository. The outcome of their study showed that MLP produced the best result on these datasets. In another work, Murty and Babu [151] considered four popular classifiers viz., NB, NN with RBF kernel (RBF-NN), MLP and DT in their work to study the performances of these classifiers on two datasets: UCI repository and MLCD. RBF-NN outperformed all other classifiers. Radhika et al. [169] used the Lung Cancer dataset from the UCI repository and SLCD. In their research, they observed that LR with a 7-fold cross validation regime and SVM with a 10-fold cross validation strategy outperformed the other classifiers on the UCI repository and SLCD dataset respectively. Recently, Patra [161] compared the performances of four different classical ML algorithms: RBF-NN, k-NN, NB and J48 on the Lung Cancer dataset of the UCI repository. The experimental outcomes confirmed that NN produces the best accuracy with a 10-fold cross validation technique. Recently, Doppalapudi et al. [53] compared the performance of 6 classifiers (3 deep learning models, namely ANN, CNN and RNN, and 3 shallow learners, namely, NB, SVM, RF) for survivability prediction of Lung Cancer prediction using SEER dataset. They experimentally showed that ANN performed the best among these classifiers while RNN and CNN models followed the ANN classifier. In another work, Gultepe [75] made a study of six different classifiers namely, k-NN, RF, NB, Logistic Regression (LR), DT, and SVM on UCI- Irvine repository and found that K-NN performed the best. The author applied the classifier to preprocessed data obtained after inputting the missing values.

In addition to these comparative studies, there are a few works that have concentrated on method building. For example, Ali and Reza [180] employed different preprocessing to filter out redundant/noisy patient data as well as attributes from the SEER dataset. In the first phase of preprocessing, the data of the patient who died due to cancer were only kept and then some irrelevant attributes like the cause of death, ID_no, age and sex were removed. Later, a correlation based FS algorithm was employed to obtain the near optimal set of attributes. After the preprocessing, the researchers selected 46 attributes from 149 attributes and 17,484 instances from 643,924 instances. They applied various ensemble methods to this optimally selected dataset. Three classifiers: J48 with the base learner, RF with Dagging and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) with AdaBoost were used and RIPPER with AdaBoost performed best.

In another work, Nasser and Abu-Naser [153] used the SLCD dataset for their work. Out of 16 attributes, in this research, the author used the first 15 attributes but excluded the chest pain attribute. ANN with 3, 1 and 2 nodes in the hidden layers was used. Salaken et al. [183] introduced a deep autoencoder based classification method. In this model, deep features were first learned and then trained an ANN with the learned features. Hinton and Salakhutdinov [87] used a DNN model with 80 and 10 features in the first and second layer respectively and the number of hidden neurons=5. They used Sigmoidal as an encoder function and got a promising result. Recently, Dey et al. [49] proposed an FS method termed LAGOA for several disease diagnosis

purposes. The selected features from LAGOA are classified using an RF classifier. The model was evaluated on the UCI-Irvine Lung Cancer dataset. In another recent work, Meleki et al. [132] used a K-NN classifier along with GA based FS approach for lung cancer prediction from the LCD dataset. Recently, Hsu et al. [95] proposed an FS scheme by combining GA with correlation based FS technique to select an optimal set of features from UCI- Irvine repository.

Some of the important findings by the researchers are listed in Table 5. Like diagnosis report based Breast Cancer detection here also FS based methods dominate. However, to the best of our knowledge, methods assisted with FS and deep learners for Lung Cancer diagnosis are still missing in the literature. However, the use of such methods could improve the diagnosis performance of the ML/DL based method significantly.

### 4.2.2 Pathological Image Based Methods

The practicalities of ML and DL based models have already been discussed in the previous sections. With the rapid advances in computational intelligence and GPUs, it has become easier for researchers to develop robust image classifiers. As a result, we find many recent methods that employed image classification protocol for Lung Cancer detection from Computed Tomography (CT) images. In the period before wider acceptability of transfer learning for learnable texture feature extraction, researchers [16, 117, 152] mostly used preprocessing to improve the visual quality of the CT images followed by handcrafted feature extraction to classify with the help of shallow learners. However, with the advent of the transfer learning concept, the research spectrum became wider. For example, Kulkarni and Panditrao [117] utilized the CT images obtained from the National Institutes of Health (NIH) or National Cancer Institute (NCI) Lung Image Database Consortium (LIDC) datasets for detecting the stages of Lung Cancer. In this work, before extracting the features, the quality of the CT images was enhanced using Median filter followed by the Gabor filter. Next, image segmentation was performed using marker based watershed segmentation [9] to extract the tumor regions. Finally, geometric features like area, perimeter and eccentricity were fed to the SVM classifier to determine the stages of cancer. Nadkarni and Borkar [152] classified CT images as cancerous (abnormal) and noncancerous (normal) using an SVM classifier placed on top of several extracted geometric features that include area, perimeter and eccentricity. Prior to the extraction of features, the median filter was applied to eradicate the noise from the CT images. Down the line, image enhancement was carried out using contrast adjustment, followed by image segmentation. Annotated CT images of Lung Cancer were collected from the Cancer Imaging Archive (CIA) repository. Two sample images taken from the CIA database are shown in Fig. 9.

Anifah et al. [16] used ANN with back-propagation and Gray Level Co-occurrence Matrix (GLCM) based features to classify 50 CT images obtained from the CIA. During prepossessing, the images were binarized and enhanced like in [152]. Finally, GLCM features were extracted from ROIs to classify the images using the designed ANN. Vas and Dessai [222] also proposed a Haralick feature assisted Lung Cancer detection technique where image segmentation was performed using morphological operation. Images were collected from the Manipal hospital and V.M.Salgaocar hospital both situated in Goa, India. The authors applied a median filter of size 3×3 to improve the input CT image quality. For ROI segmentation a four steps method was employed. The steps are 1) complementing the image and opening the image with periodic-line structuring elements, 2) filtering out the lungs by using maximum area, 3) performing a close operation by disk structuring element to procure lung mask and 4) superimposing the lung mask on the original image. 7 Haralick features (second order statistical features) were extracted by applying Haar wavelets and creating GLCM from the wavelet transformed images. Recently, Shakeel et al. [198] used deep learning instantaneously trained neural network (DITNN) classifier on top of several statistical and spectral features designed by Sridhar et al. [209] aiming at designing a Lung Cancer detection model using CT images. The main contribution of the authors is proposing an improved profuse clustering technique (IPCT) that segments the affected regions from the CT image efficiently and thus improves the overall detection result. Alzubaidi et al. [13] performed an experimental comparison between global and local hand-crafted features used for Breast Cancer detection from CT images. For the comparative study, 100 CT images were collected from different sources while 10 different feature extraction processes, namely intensity histogram, HOG, Gabor Haar wavelet, etc. This comparative study confirms that the use of local Gabor features with the SVM classifier performed the best. Recently, Akter et al. [7] designed a fuzzy rule based image segmentation technique to extract suspicious nodules from a CT image. Next, they prepared shape (2D and 3D) and texture (2D) based features from the extracted nodules and classified them using a neuro fuzzy classifier. In another work, Zhou et al. [236] first employed a region growing mechanism to segment suspicious regions from a CT image and then a CNN model to classify the segmented regions to decide the presence of Lung Cancer.

In the era of DL with the concept of transfer learning, we observed some works that utilized the same in their work. For example, Jakimovski and Davcev [104] gathered CT images of 70 subjects from a local medical hospital and labelled them from oncology specialists. Two piles were created, where 58 subjects had cancer, and 12 were diagnosed cancer-free. The piles were split into training (90% of the images) and testing sets. After binarization, images were

Table 6 Performance comparison of recent prediction models on Lung Cancer Datasets

Work Ref.	Classifier	Performance metrics				
		ACC	P	R	F1	S
CIA						
Tekade and Rajeswari [218]	VGG	0.95	–	–	–	–
Anifah et al. [16]	DNN	0.80	–	–	–	–
Alam et al. [10]	SVM	0.97	–	–	–	–
Shakeel et al. [198]	DITNN	0.95	0.94	0.97	0.95	0.97
Zhang and Kong [231]	MSDLF	0.99	–	–	–	–
Mics. Datasets						
Jakimovski and Davcev [104]	DNN	0.75	–	–	–	–
Taher et al. [216]	Rule-based classifier	0.93	0.95	0.94	–	0.91
Vas and Dessai [222]	DNN	0.92	–	0.88	–	0.97
Alzubaidi et al. [13]	SVM	0.97	–	0.96	–	0.97
Akter et al. [7]	Neuro-Fuzzy	0.90	0.86	–	–	0.81

Fig. 9 Examples of Lung CT images from CIA

(a) Normal　　　　　　　　　　(b) Lung cancer
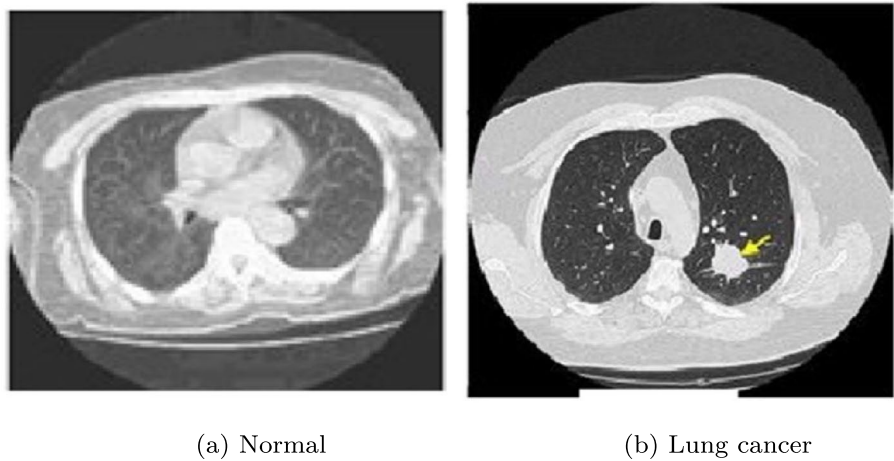
fed into the DNN for classification. In another work, Taher et al. [216] came up with a rule based method that analyzed the sputum samples obtained from the Tokyo Center of Lung Cancer in Japan. The features extracted from the nucleus region included nucleus-cytoplasm ratio, perimeter, density, curvature, circularity and eigen ratio. Diagnosis rules for each of the features were derived. 100 sputum color images were used to assess the rule based method. Finally, classification was done using a rule based method. While Kulkarni and Panditrao [117] used classical ML in their work, Tekade and Rajeswari [218] proposed a 3D multipath VGG-like CNN architecture. In the preprocessing stage like image binarization, morphological operations were applied. After preprocessing, U-Net architecture was used for the segmentation of lung nodules, which were fed to the CNN model for detection. Finally, the features were fed to an ANN for classification needs. Since identified lung nodules greatly help in risk assessment of Lung Cancer, Zhang and Kong [231] designed a Multi-Scene Deep Learning Framework (MSDLF) to detect Lung Nodule from Lung CT images.

The detected nodules were classified using a 4-channel CNN architecture. The summary of performances of all the methods studied here is provided in Table 6.

4.3 Leukemia Prediction Methods

Considering the mortality rate of Leukemia, researchers have made significant inroads into the early detection of Leukemia using both gene expression and pathological image. All genes present in gene expression data are not responsible for Leukemia, rather there are a few which known as biomarkers. Thus, selecting the responsible genes was vastly explored in literature. Old days method used filter methods while the current trend is to use different hybrid FS mechanisms for the said purpose. In the case of image based methods, a shift from the traditional ML approach to the DL approach is observed. In the following subsections, we

Table 7 Performance comparison of recent prediction models on Leukemia Datasets. All the methods were evaluated on the Dataset provided by Golub et al. [70]

Work Ref.	Classifier	Performance metrics			
		ACC	R	F1	S
Hsieh et al. [94]	IG-SVM	0.98	–	–	–
Begum et al. [27]	CBFS-Poly SVM	0.97	–	–	–
Kavitha et al. [114]	FCBF-SVM RFE	0.93	0.95	–	0.98
Dwivedi [55]	ANN	0.98	1.00	–	0.93
Gao and Liu [63]	PSOFOA-ANNFIS	1.00	–	–	–
Santhakumar and Logeswari [187]	ALO-SVM	0.91	0.95	0.93	0.95
Santhakumar and Logeswari [188]	ALO-ACO-SVM	0.95	0.95	0.96	0.98
Baldomero-Naranjo et al. [24]	Modified SVM	0.95	–	–	–
Mandal et al. [133]	Tri-state FS	1.00	–	–	–

describe the methods that followed either gene expression or cell image based diagnosis.

4.3.1 Gene Expression Based Methods

In this study, the classification of Leukemia into Acute Lymphoblastic Leukemia (ALL) or Acute Myeloid Leukemia (AML) is considered. The most common dataset which has been used is the dataset[16] provided by Golub et al. [70]. This is a gene expression dataset. It consists of 72 Leukemia patients (47 in the ALL category and 25 in the AML category) and each data consists of 7129 gene expression measurements. The researchers aimed to come up with a near optimal set of genes that are responsible for Leukemia disease prior to designing a classifier based detection technique. For example, Hsieh et al. [94] used Information Gain (IG) to select important genes that play a significant role in identifying the presence of Leukemia. The authors selected 150 genes out of 7128 genes using the IG. These 150 genes were later used to classify genes using SVM with the RBF kernel. In another work, Begum et al. [27] performed Consistency based FS (CBFS) to select optimal bio-markers (i.e., genes) that helped distinguish ALL from AML with SVM (with polynomial kernel) classifier. Contrary to the above methods, Kavitha et al. [114] used Fast Co-relation based Filter-Solution (FCBF) as feature selector and SVM with Recursive Feature Elimination (RFE) kernel as the classifier. In FCBF, first Symmetrical Uncertainty (SU) was computed for every feature and then the correlated features were deleted based on a preset threshold value.

In another work, Dwivedi [55] made a comparison study among many classifiers like ANN, SVM, LR, k-NN and classification tree. For cross-validation, researchers used to leave one out cross validation. The best result came out for ANN with the leave one out cross validation technique. Recently, Santhakumar and Logeswari [187] used Ant Lion Optimizer

(ALO), designed by Mirjalili [139]. After FS, the SVM classifier was applied to the selected features. In another work [188], these authors proposed a hybridized ALO with Ant Colony Optimization (ACO) technique to improve their previous performance. Though these two methods are new approaches, their performance was not so good compared to other works on the said dataset. In another work, Gao and Liu [63] gained the highest accuracy with the help of an extension SVM classifier termed Least Square SVM (LSSVM) where the authors first normalized the data and then classified using LSSVM. Next, the F-statistic method was used as a filter method to lessen the complexity of SVM-RFE. Both PSO and FOA tend to get stuck in local optima. Hence, FOA was used first and then PSO was used to reach the global optimum if FOA stuck to the local optimum. Finally, LSSVM is used on the select features and they got an extraordinary result of 100% with 4 features.

Recently, Baldomero-Naranjo et al. [24] designed a modified SVM classifier that deals with outliers detection and FS simultaneously. They considered ramp loss margin error [31] in the newly designed SVM model to mitigate the influence of outlier on the classifier and a budget constraint approach similar to [123] to restrict the number of the selected features. In another work, Baliarsing et al. [25] designed a hybrid FS method that combines SA with Rao Algorithm (RA) (termed as SARA). The SA based local search helps in improving the exploration capability of the RA method. They also used Log sigmoidal function as a transfer function to convert the continuous domain SARA into the discrete domain. Mandal et al. [133] proposed a tri-state wrapper-filter FS method and evaluated the performance on the dataset provided by Golub et al. [70]. In the first stage, four filter methods, namely Mutual Information, ReliefF, Chi-Square, and Xvariance were considered to design an ensemble FS that reduces some irrelevant features from high-dimensional features. In the next stage, the highly correlated features

---

[16] https://web.stanford.edu/~hastie/CASI_files/DATA/leukemia.html

Fig. 10 Samples taken from ALL_IDB images

(a) Leukemia cell

(b) Leukemia cell

(c) Normal Cell

(d) Normal Cell
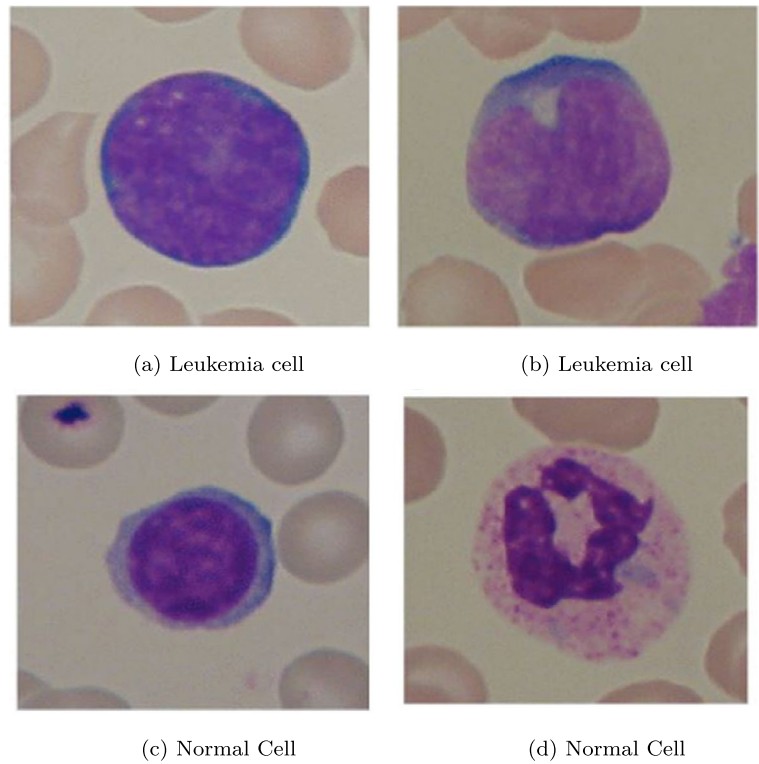
Fig. 11 Example of a peripheral blood smear having (a) lymphoblast and (b) lymphocyte (ALL_IDB Dataset)
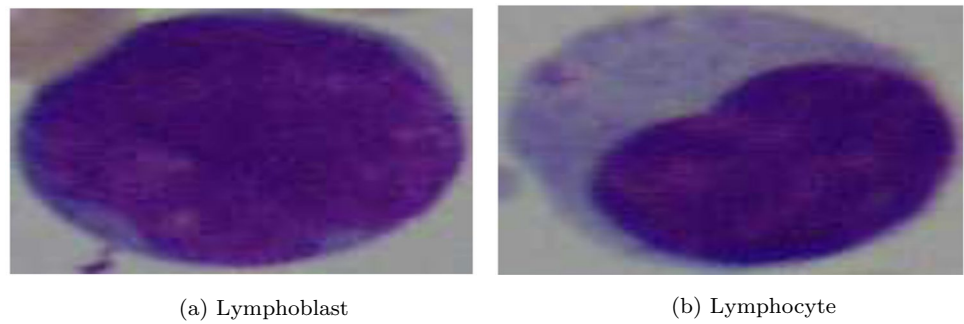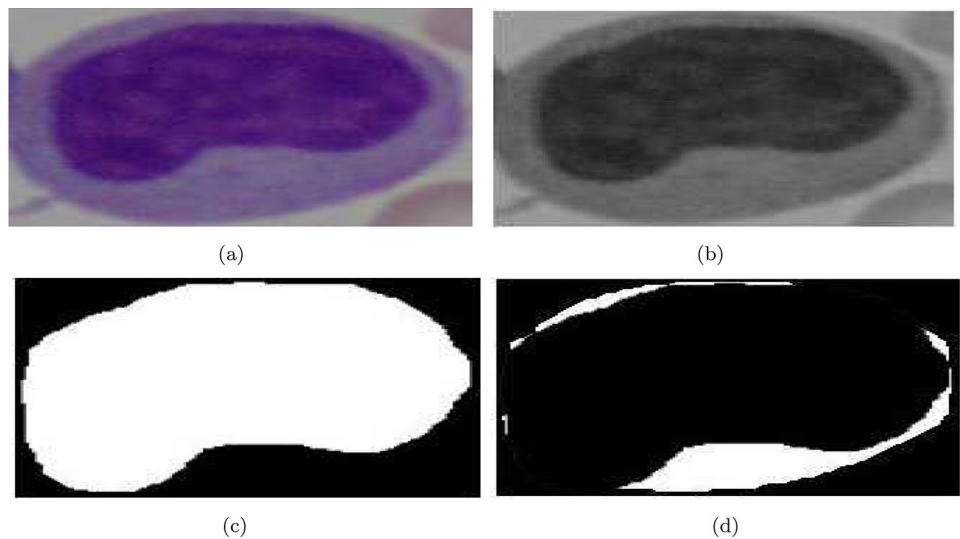

(a) Lymphoblast

(b) Lymphocyte

Fig. 12 Separation of nucleus and cytoplasm exercised by Mishra et al. [143] in their work

(a)

(b)

(c)

(d)

were filtered out by employing the Pearson correlation based method. Finally, the authors employed WOA to select the final set of biomarkers. Table 7 lists the performance of various research efforts discussed here.

4.3.2 Pathological Image Based Methods

Since Leukemia disease usually involves white blood cells and thus the nature of these cells helps in screening it. As a result, the pathological image based Leukemia detection algorithm relied on understanding the nature/proportion of blood cells. For this category of Leukemia classification methods considered here have used the Acute Lymphoblastic Leukemia Image Database (ALL_IDB[17]). All images are in JPG format with 24-bit color depth and dimension 2592×1944. Some sample images of ALL_IDB dataset are shown in Figs. 10 and 11. The dataset has two parts: the ALL_IDB 1 and ALL_IDB 2. The ALL_IDB 1 dataset with 108 samples is used primarily for differentiating patients from non-patients. It contains about 39000 blood cells, where the lymphocytes have been labelled by expert oncologists. The ALL_IDB 2 dataset is primarily used to classify different types of Leukemia. It contains a collection of cropped areas of interest of normal and blast cells (see Fig. 11) that belong to the ALL_IDB1 dataset (see Fig. 10). ALL_IDB2 images have similar grey level properties to the images of the ALL_IDB1, except for the image dimensions. Hereafter we are discussing some significant works in Leukemia disease detection.

Abdeldaim et al. [2] utilized the ALL_IDB 2 dataset to detect different types of Leukemia disease. In this work, the images were first converted into a CMYK color model and then used histogram based threshold calculation method to separate the lymphocyte cells from the non-lymphocyte cells where roundness and solidity characteristics of the cells were considered. After this, 30 shape-based features, 15 color based features and 84 texture based features were extracted from detected lymphocyte cells and stacked to classify the cells using k-NN, SVM (with RBF, Polynomial and Linear kernels), DT, NB classifiers. k-NN outperformed the other classifiers. Mishra et al. [142–144] proposed a series of solutions in the said field with varying features and classifiers. In all these works, they utilized the Weiner filter followed by histogram equalization based contrast enhancement on the images from ALL_IDB 2 dataset. Next, grouped leukocytes were segmented using marker based watershed segmentation algorithm [125]. Some such segmented outputs are shown in Fig. 12).

In the work, proposed by Mishra et al. [142], Discrete Cosine Transform (DCT) features were extracted from

isolated leukocyte cells and classified using SVM, NB, k-NN and NN with Back Propagation (BPNN). It was found that the first 50 DCT coefficients with the SVM classifier performed the best. Next year, in the work [143], they first extracted GLCM features from segmented cells and then they employed probabilistic PCA for feature dimension reduction. Finally, they used RF, k-NN, SVM and BPNN classifiers on dimension reduced features for both nuclei as well as cytoplasm detection from which it was inferred that nucleus features were more suitable for accurate detection of Leukemia. In another work [144], the authors performed texture based cell classification for predicting Leukemia. Texture based features were extracted using Discrete Orthonormal S-Transform (DOST) from each segmented region. This was followed by dimensionality reduction using LDA. For classification, the authors tried AdaBoost, SVM, BPNN, RF and k-NN out of which AdaBoost yielded the best result.

Similar to Mishra et al. [142–144], Shafique et al. [196, 197] also proposed two different solutions for classifying Leukemia but they used state-of-the-art transfer learning models for feature extraction purposes in [196] while in [197] they preferred hand-crafted color and shape based features and SVM classifier. In transfer learning, the last layer was changed first to detect white blood cells in the images and then to 4 output channels to categorize the subtypes of Leukemia. This process was done for RGB, HSV, YCbCr and HCbCr image types and used data augmentation during training. The results depicted that the best classification was observed in the case of RGB. In another work [197], the authors performed white blood cell extraction using a series of preprocessing steps (conversion to CMYK and histogram equalization based contrast enhancement), white blood cells suppression using Zack's algorithm, segmentation of image using watershed algorithm prior to classification. Whereas, Rahman and Hasan [170] detected white blood cells in four major steps which involved the conversion of the image from RGB to HSV, histogram equalization based contrast stretching, background removal and detection using the watershed algorithm. The authors then performed image cleaning to remove leukocytes situated at the edges and nucleus and cytoplasm separation using the bounding box technique. After this, the authors performed feature extraction to obtain 70 features: 16 morphological features, 36 texture features and 18 color features. The classification was performed using an ensemble of classifiers having SVM, DT and k-NN as base classifiers.

In contrast to the above methods, Rawat et al. [176] used the French American British (FAB) classification system for ALL_IDB 2 dataset. According to the FAB, there are three types of Leukemia in the ALL_IDB 2 dataset (say, L1, L2 and L3). Based on this, the authors designed a hierarchical classification model. In the first stage, normal cells and cancerous cells were classified. In the next stage, from

---

Table 8 Performance comparison of recent prediction models on ALL_IDB dataset

Work Ref.	Classifier	Performance metrics				
		ACC	P	R	F1	S
ALL_IDB 1						
Mishra et al. [142]	SVM	0.99	0.99	1.00	0.99	0.99
Shafique et al. [197]	SVM	0.93	–	–	–	–
Rahman and Hasan [170]	SVM+DT+k-NN	0.93	–	0.96	–	–
Mohammed et al. [145]	SVM	0.97	–	–	–	–
ALL_IDB 2						
Abdeldaim et al. [2]	k-NN	0.96	–	–	–	–
Mishra et al. [142]	SVM	0.89	–	0.84	–	0.94
Shafique et al. [196]	DNN	0.99	–	–	–	–
Rawat et al. [176]	ANFIS	0.96	–	–	–	–
Mishra et al. [143]	RF	0.99	0.97	1.00	0.98	0.98
Sahlol et al. [181]	SVM	0.98	1.00	0.97	0.985	0.97

Table 9 Performance comparison of some prediction models on Cleveland dataset

Work Ref.	Classifier	Performance metrics					
		ACC	P	R	F1	S	ERR
Single classifier based methods							
Al-Milli [8]	BPNN	0.92	–	–	–	–	–
Sonawane and Patil [208]	BPNN	0.985	–	0.98	–	0.95	0.02
Gavhane et al. [64]	MLP	–	0.91	0.89	–	–	–
Karayılan and Kılıç [113]	BPNN	0.96	0.95	0.95	–	–	–
Ismaeel et al. [100]	ELM-NN	0.87	–	–	–	–	–
Medhekar et al. [136]	NB	0.90	–	–	–	–	–
Jabbar et al. [101]	NB	0.86	–	–	–	–	–
Patel et al. [158]	J48	–	–	–	–	–	0.15
Dwivedi [56]	LR	0.85	0.85	0.89	–	0.81	–
Sen [194]	SVM	0.84	–	–	–	–	–
Shao et al. [200]	MARS-LR	0.84	–	–	–	–	–
Jayaraman and Sultana [106]	PBAMNN	1.00	–	1.00	–	1.00	–
Gokulnath and Shantharajah [69]	GA-SVM	0.88	–	–	–	–	–
Feshki and Shijani [60]	PSO-FFBP	0.92	0.92	0.93	–	–	–
Harimoorthy and Thangavelu [81]	Improved SVM	0.90	0.81	0.97	–	0.87	0.10
Ensemble based methods							
Das et al. [48]	Multi-NN	0.89	–	–	–	–	–
Miao et al. [138]	Boosting	0.97	0.97	1.00	0.98	0.60	0.03
Mohan et al. [146]	HRFLM	0.88	0.90	0.93	0.90	0.83	0.12
Maji and Arora [130]	DT-NN	0.78	–	0.78	–	–	0.23
Amin et al. [14]	Average voting	0.87	–	–	–	–	–

amongst the cancerous cells, L1 and non-L1 cells were classified. Similarly, L2 and L3 were classified in the third stage. The authors first extracted 11 geometrical, 15 chromatic and 45 statistical texture features from preprocessed images followed by PCA to obtain reduced feature vectors. In each stage of the hierarchical classification model k-NN, Probabilistic NN (PNN), SVM, Smooth SVM (SSVM) and ANN based Fuzzy Inference System (ANFIS) were applied. The experimental outcome showed that SVM performed best

based on classification accuracy. Mohammed et al. [145] proposed a two-stage method where first white blood cells were segmented using a series of preprocessing techniques like median filtering, histogram equalization and hard thresholding and then features like shape, statistical, texture and DCT features were extracted from the segmented images to classify the cells as abnormal or healthy. For classification, the authors used an SVM classifier. Recently, Sahlol et al. [181] proposed a wrapper FS method termed

as statistically enhanced Salp Swarm Algorithm (SESSA), where SVM was used as a classifier, to select deep features, extracted using VGG architecture, for the classification of pre-segmented white blood cells. Recently, Sumi et al. [213] proposed a self-made CNN-architecture to detect presence of Leukemia. To train the model, they generated synthetic samples through a data augmentation process. In another work, Khandekar et al. [115] utilized an object detection algorithm [147] known as the You Only Look Once (YOLO) algorithm (version 4) to extract and classify ROIs simultaneously.

Performance comparison of all the methods discussed above has been listed in the Table 8. It is to be noted that these performances were not obtained following uniform experimental setups and thus are not directly comparable. However, after closely analyzing the results, it is observed that hand-crafted features provided better performance over deep feature techniques [181, 196]. The reason might be that the CNN models used in literature are of larger depth and thus tend to extract texture features which in turn failed to catch the shape information. Hence, in the future, the use of some better CNN models that can highlight shape and texture information may be useful. Even the use of FS and classifier ensemble techniques are not been largely explored to date.

5 Heart Disease Prediction Methods

One of the common datasets which have been used by many researchers is the Heart disease dataset from the UCI repository[18]. This is a diagnostic report based dataset that contains 76 attributes in total (1 class attribute and 75 predictive attributes). However, most of the published works used a subset of 14 attributes to conduct their experiment. In particular, the Cleveland dataset is the one that has been very often used by researchers for Heart disease prediction. The numbers 0 and 4 indicate no Heart Disease and end of life respectively while the numbers 1, 2 and 3 indicate the corresponding attack count. The dataset consists of 303 subjects and their corresponding details. In this study, we categorize the Heart Disease detection techniques into two major categories based on whether a method employs a single classifier or an ensemble of classifiers in the classification system. Table 9 lists the performance of various research efforts on the UCI heart dataset discussed here.

5.1 Single Classifier Based Methods

In this section, we have discussed the works which utilized a single classifier for Heart Disease classification. The single

classifier based systems include the use of variants of NN models, FS technique prior to classification, missing value imputation, and comparative study to come up with a better model. Since all the diagnostic attributes present in the dataset might not be essential always while screening the Heart disease, hence many authors have utilized FS algorithms to build a state-of-the-art prediction model. Relatively older methods that utilized some form of feature reduction mechanism mostly used the filter FS techniques. Next, we observe a shift from filter to wrapper FS methods and at present hybrid FS techniques are being proposed. Many works also consider this UCI Heart disease dataset to show the effectiveness of newly developed FS models.

In the literature, there are a large number of methods that used NN based models. A number of works like Al-Milli [8], Sonawane and Patil [208] and Gavhane et al. [64] cited the state-of-the-art results on Heart Disease classification problem by tuning the parameters or increasing the number of hidden layers in an MLP with backpropagation algorithm. For example, in [8] though the iteration size is fixed there were 3 hidden layers with 8, 5, and 2 neurons. Whereas in some experiments only one hidden layer with a different number of neurons produced comparatively better results. For example, the numbers of neurons in hidden layer were 20, 100 and 8 in Sonawane and Patil [208], Gavhane et al. [64] and Karayılan and Kılıç [113] respectively.

Jabbar et al. [101] and Kanchan et al. [110] established the benefits of FS for classifying Heart Disease. Jabbar et al. [101] utilized the One-R FS technique, proposed by Holte [88] produced better results (combining with NB) compares to Chi-square, Gain Ratio and ReliefF. Shao et al. [200] and Feshki and Shijani [60] employed FS strategy prior to actual classification using LR and Fed Forward BP (FFBP) NN respectively. Shao et al. [200] used a rough set while in [60], PSO was employed for FS purpose by the authors. Using PSO Al-Milli [8] showed the importance of good FS technique as PSO-FFBP produced better results only after utilizing it (utilized multiple hidden layers) in terms of accuracy with a single classifier. In [186], the authors selected the best feature using the ICRF-LCFS algorithm, after that CNN was applied with temporal features with dense layers and activation function as RELU.

In another work, Jayaraman and Sultana [106] used Artificial Gravitational Cuckoo Search Algorithm (AGCS) for FS. As a classifier authors utilized the Particle Bee Optimized associative Memory NN (PBAMNN), where the features were examined in the layers of the NN and this examination was done using appropriate weights and sigmoid as the activation function. Kanchan et al. [110] and Karayılan and Kılıç [113] employed PCA for dimensionality reduction. The latter reduced the number of attributes from 13 to 8 and the number of neuron in the hidden layer was decided by an iterative process. In [69], the authors utilized GA based

---

Table 10 Performance comparison of recent prediction models on PIMA Indians Diabetes dataset. Classifiers with '*' represent the classifiers with the highest ACC scores in the respective paper

Work Ref.	Classifier	Performance metrics					
		ACC	P	R	F1	S	AUC
Patil et al. [160]	DT	–	–	0.79	–	0.93	0.95
NirmalaDevi et al. [154]	k-NN	0.97	–	0.97	–	0.97	–
Chen et al. [39]	DT	0.90	–	0.87	–	0.91	–
Wu et al. [227]	LR	0.95	0.95	0.95	–	–	0.98
Sisodia and Sisodia [206]	NB	0.76	0.76	0.76	0.76	–	0.82
Dutta et al. [54]	RF	–	–	0.84	0.84	–	–
Zhu et al. [237]	LR*	0.97	0.97	0.97	–	–	–
Daanouni et al. [46]	DNN	0.90	–	–	–	–	–
Rakshit et al. [173]	ANN	0.83	–	–	–	–	–
Ashiquzzaman et al. [19]	MLP	0.88	–	–	–	–	–
Saji and Balachandran [182]	MLP	0.70	–	–	–	–	–
Jahangir et al. [102]	MLP	0.89	0.85	0.88	–	–	–
Hasan et al. [82]	AB+XGB	–	–	0.79	–	0.93	0.95
Kannadasan et al. [111]	Encoder part of autoencoder	0.86	0.90	0.87	0.89	0.83	–
Vaishali et al. [221]	MOE NSGA II	0.83	–	–	–	–	–
Mansourypoor and Asadi [135]	RLEFRBS*	0.84	–	–	–	–	–
Harimoorthy and Thangavelu [81]	Improved SVM	0.99	1.00	0.95	–	1.00	–

approach for selecting the optimal features. After FS, out of 13 features, only 7 features were selected and then SVM was used for classification. Recently, Harimoorthy and Thangavelu [81] selected the features using Chi-square based filter method. However, their major contribution was that they improved the radial basis kernel of the SVM model where they iteratively decreased the margin size i.e., increased the cost during model training. Doing so they reduced the overall classification error.

Missing values can have a detrimental effect on the quality of data. To ameliorate this, Choudhury and Pal [43] introduced a novel method of missing value imputation using autoencoder NN. The dataset was trained on multiple datasets like UCI Cleveland dataset without any missing values for the attribute. The trained autoencoder was then used to predict missing values. As an initial guess of the missing value, the nearest neighbor rule was used and then refined by minimizing the reconstruction error. This was based on the hypothesis that a good choice for a missing value would be the one which can reconstruct itself from the autoencoder. The classifiers used for the imputation were SVM, LR, NB, MLP, PRZ, k-NN, CART, and PNN. The authors comprehensively established the superiority of their imputation strategy especially when there is a high number of missing values. On the other hand, Ismaeel et al. [100] used fixed hidden layer neurons (100) and Extreme Learning Machine (ELM) [97] to get rid of the slow training capability of BPNN algorithms as ELM is a one-time pass process thought it would be fast but it failed to acquire accuracy like the work proposed by Sonawane and Patil [208].

Apart from the NN based classifiers, some researchers have used other classifiers like NB [136], J48 [158], LR [56] and SVM [194]. In the work by Medhekar et al. [136], the NB classifier was trained using 240, 276 and 290 samples out of 303 as training samples where the set with 240 samples outperforms all others. Sen [194] experimentally showed that SVM performs better than DT, NB and k-NN. Whereas in the work by Dwivedi [56], it was shown that LR performs better than SVM, DR, NB and k-NN.

5.2 Ensemble Based Methods

In the literature, some Heart Disease prediction systems empowered by several ensemble techniques are also reported which make use of various classifiers to build an integrated classification system. Das et al. [48] proposed an NN ensemble classifier in their research. They first rejected the attributes which were not relevant to Heart Disease classification. This was followed by dividing the dataset into two or more mutually exclusive subsets first and then NN was applied to each dataset to form an ensemble of NNs that were trained with different configurations. Miao et al. [138] proposed an ensemble method based on AdaBoost and weighted majority vote. The researchers found that the boosting methods were very powerful compared to other ensemble methods.

Recently, Maji and Arora [130] proposed an ensemble technique that ensembles DT (C4.5) and NN classifiers named Hybrid DT. In another work, Mohan et al. [146] introduced a method named hybrid RF with a linear model (HRFLM) to set new benchmark recognition accuracy on

the UCI Heart Disease dataset. In this model, a DT based partition was applied first, which in turn generated leaf nodes with constraints and then the leaf nodes were pruned using an entropy score. Next, using the constraint of the leaf node, the dataset was split into several datasets (same as the number of leaf nodes). Each dataset was classified separately using the linear method and RF. In contrast to the mentioned methods, Amin et al. [14] proposed a new idea for FS, where they applied a brute force approach and all combinations of features having at least 3 features were tested. In their experimentation, with 13 input attributes, a total of 8100 combinations were selected and tested. After the FS, a model with popular 7 different classifiers (NB, SVM, k-NN, DT, LR, NN, and Vote (average voting of NB and LR)) was created and the dataset was fed into that. In this experiment, with 9 features, Vote method produced the maximum classification accuracy.

6 Diabetes Prediction Methods

In this section, we chronicle some of the research efforts undertaken by the researchers to classify patients as diabetic or non-diabetic. The works on Diabetes prediction considered in the current survey used diagnosis report based classification utilizing the PIMA Indians Diabetes dataset[19] prepared by the National Institute of Diabetes and Digestive and Kidney Diseases. All the subjects considered in this dataset were female and above the age of 21. The dataset has 8 features namely, the number of pregnancies, body mass index (BMI), insulin level, glucose, blood pressure (BP), skin thickness, Diabetes pedigree function and age, and a single target variable: Outcome that uses 0 and 1 for non-diabetic and diabetic cases respectively. The research efforts on PIMA Indians Diabetes dataset in the past decade have been enumerated in the Table 10. Some of the ideas presented in these works are as follows.

One of the most cited papers in the field of Diabetes prediction is by Patil et al. [160]. The authors introduced a novel hybrid prediction model (HPM) for Diabetes which has been extensively used and upgraded by other researchers to have a better performing model. The method performed normalization and missing value deletion in the preprocessing stage followed by data reduction using a k-means clustering algorithm. k-means clustering was used to classify the training data into two clusters: diabetic and non-diabetic. In this way, the samples were demarcated into two clusters, out of which the wrongly classified instances were omitted from training followed by classification using DT (J48 algorithm). Later, NirmalaDevi et al. [154] used an amalgam of

k-means clustering and k-NN classifier for the prediction of Diabetes. In this method, prepossessing was first done by removal of inconsistent values followed by k-means clustering to remove the misclassified data from the training set as suggested by Patil et al. [160]. However, unlike Patil et al. [160], k-NN was used as the classifier. Though the usage of an HPM was novel, the dataset shrinks further due to the removal of misclassified data.

Significant research developments with greater focus on the data preprocessing like missing value imputation [39, 43, 227], class imbalance handling [54, 202, 203], outlier removal [102], and dimensionality reduction [204, 221] have taken place over the years. We have described a few important works that dealt with such processing to improve prediction performance. Chen et al. [39] improved upon the HPM model of their predecessors by improvising a new missing value imputation method. The authors imputed missing values using the mean of the remaining values of the respective features and used SVM as the classifier. The work by Wu et al. [227] was also along similar lines, except that they used LR as the classifier. Choudhury and Pal [43] introduced a novel method of missing value imputation using autoencoder NN. The method autoencoder model trained was trained on the dataset without any missing values. Next, the trained autoencoder was used to predict missing values in other records. As an initial guess for the missing value, the nearest neighbor rule was used and then refined by minimizing the reconstruction error. A variety of missing value imputation techniques were considered by Purwar et al. [168]. The authors implemented an HPM using MLP. Missing values were imputed using 11 different techniques like case deletion, most common method, concept common method, k-NN, weighted k-NN, k-means clustering, Fuzzy k-means clustering, SVMs imputation (SVMI), singular value decomposition based imputation (SVDI), local least square imputation and matrix factorization, each of which were classified using k-means clustering. The dataset with the least misclassification rate was selected. This new imputed dataset without the misclassified data was classified using MLP to predict Diabetes. However, the dataset is imbalanced, misclassification rate is not an appropriate metric to judge the missing value imputation technique.

To deal with the problem of class imbalance, Dutta et al. [54] used oversampling of the minority class samples. Various ML algorithms like SVM, RF and LR were used for classification and RF yielded the best result. However, oversampling leads to the synthetic generation of data which is not entirely reliable for sensitive domains like disease prediction. To deal with the missing values, Harimoorthy and Thangavelu [81] removed missing values by ignoring the missing fields. After using the X-square filter FS, the authors detected the presence of disease using the modified redial basis kernel of SVM. It is to be noted that recently,

---

Shaw et al. [202, 203] proposed two new methods (ensemble method [203] and RTPSO for majority sample selection [202]) for better handling imbalance class problem in a dataset.

Another 'impurity' in real world data is the existence of outliers [34]. Significant research efforts have been devoted to eliminating this problem in the domain of diagnostic report based disease prediction. For instance, Jahangir et al. [102] designed an outlier detection technique to preprocess the data. The outliers in the dataset were detected using the enhanced class outlier factor (ECOF) based method that improved the supervised class outlier factor (COF) based method. COF for a particular instance is measured using three different scores: the probability of being part of the class under consideration (i.e., $\frac{n}{k}$, where n represents the number of samples from its k neighbours belong to the same class), the deviation of from the respective instances of the same class and the average distance from its neighbors. However, the tuning parameters COF was replaced by a normalization factor in ECOF based method. After this, the authors used an auto tunable NN that adjusts the learning rate and the number of hidden units for classification. An inappropriate outlier deletion method may lead to the removal of a large amount of training data which may cause underfitting/overfitting of NN based classifiers. Hence, a better outlier detection method like [33] could be used.

Zhu et al. [237] and Wu et al. [227] converted the number of pregnancies into a categorical feature to distinguish between zero and non-zero pregnancies as a part of data preprocessing but followed different classification strategies. Zhu et al. [237] first applied PCA on preprocessed data to reduce the number of features and then they used k-means clustering to delete the misclassified instances from the training set. However, Wu et al. [227] used k-means clustering based training data deletion on the original data samples. Wu et al. [227] used the basic version of LR while Zhu et al. [237] used improved LR as the underlying classifier. Here, it is to be noted that by clubbing all women with a history of pregnancy into a single bin, the authors disregarded the number of pregnancies. However, according to medical research [127], women with at least four pregnancies were 28% more likely to develop Diabetes compared to women who reported two or three pregnancies. Hence, converting the number of pregnancies into a binary variable may lead to medical inconsistencies.

Apart from data preprocessing, another familiar problem that researchers face is the problem of FS. Quite often, the datasets are encumbered with irrelevant information which needlessly increases the complexity of the model. FS has been widely explored for a long time starting with Ilango and Ramaraj [99] who used an HPM approach. In this work, a hybrid FS technique that first ranked the features based on F-score followed by optimum feature subset selection having the least clustering error was used. SVM was used for classification. It is noteworthy that F-score does not provide any mutual information among the features. Hence, it may lead to the selection of multiple irrelevant features which convey the same information. In another work, Vaishali et al. [221] used GA for FS which reduced the feature dimension to half its actual size. After the FS, two types of multi-objective evolutionary (MOE) fuzzy rule based classification algorithms namely, Non-dominated Sorting GA (NSGA II) & elitist Pareto-based multi-objective evolutionary algorithm (ENORA) were employed. MOE NSGA II fuzzy algorithm yielded a better result. In the same line, Shen et al. [204] applied FOA-SVM technique. This technique can select optimum features with an iterative method till the number of iterations matches the maximum iteration present. The FOA-SVM technique produced very good results compared to methods like PSO-SVM and GA-SVM.

Fuzzy systems were also employed by Mansourypoor and Asadi [135] where the authors designed a Reinforcement Learning-based Evolutionary Fuzzy Rule-Based System (RLEFRBS). The proposed model involved the building of a Rule Base (RB) followed by rule optimization. The initial RB was constructed from numerical data out of which redundant rules were eliminated based on the confidence measure. This was followed by rules used for pruning. Finally, GA was employed to select the appropriate subset of rules. Usage of fuzzy systems ensures high interpretable models which makes such models acceptable in the research community. Both the problems of data preprocessing and FS are dealt with elegantly by Hasan et al. [82]. The authors studied the performances of two ensemble learners (viz., XGBoost, and AdaBoost) and MLP in three different scenarios. In the first case, preprocessing techniques like Outlier Elimination (OE), Missing Value Substitution (MVS), and standardization were applied with dimensionality reduction techniques like PCA, ICA and FS using correlation. In the second method, MLP was designed after applying OE and MVS. In the third method, OE + MVS with AdaBoost + XGBoost was found to be the best ensemble technique. It is observed that the first method had the best specificity while the other method yielded the best values in sensitivity and AUC score.

Apart from these, we find extensive usage of MLP in Diabetes prediction. Saji and Balachandran [182] and Ashiquzzaman et al. [19] made significant contributions to this domain. Saji and Balachandran [182] compared the outcomes of MLPs with various learning algorithms like Levenberg-Marquardt, BFGS Quasi-Newton, Resilient BP, scaled conjugate gradient, conjugate gradient with Powell/Beale restarts, Fletcher-Powell conjugate gradient, Polak-Ribiere conjugate gradient, one step secant and variable learning rate backpropagation. Levenberg-Marquardt yielded the best results among all. Ashiquzzaman et al. [19] experimented

Table 11 Performance comparison of recent prediction models on UCI CKD dataset. Classifiers with '*' represent the classifiers with the highest ACC scores in the respective paper

Work Ref.	Classifier	Performance metrics					
		ACC	P	R	F1	S	AUC
Pujianto et al. [167]	SVM	1.00	–	–	–	–	–
Amirgaliyev et al. [15]	SVM	0.94	–	0.94	–	0.95	–
Chetty et al. [41]	k-NN*	1.00	–	–	–	–	–
Charleonnan et al. [35]	SVM	0.98	–	0.99	–	0.98	–
Wibawa et al. [226]	AB	0.98	0.98	0.98	0.98	–	–
Sedighi et al. [190]	NB	0.97	–	–	–	–	–
Sisodia and Verma [207]	RF	1.00	1.00	1.00	1.00	–	1.00
Avci et al. [21]	DT	0.99	0.98	0.99	0.98	–	–
Almansour et al. [12]	ANN	0.99	–	–	–	–	–
Wibawa et al. [225]	SVM	–	–	0.99	–	1.00	–
Aljaaf et al. [11]	MLP	0.98	–	–	0.98	–	0.99
Gunarathne et al. [76]	RF	0.99	–	–	–	–	–
Guia et al. [74]	k-NN	–	–	–	0.99	–	–
Johari et al. [108]	ANN	1.00	1.00	1.00	–	–	–
Polat et al. [164]	SVM	0.98	0.98	0.98	–	–	0.98
Harimoorthy and Thangavelu [81]	Improved SVM	0.98	0.95	1.00	–	0.99	–

with MLPs with different dropout schemes. The model used 2 hidden layers, one with 64 neurons and the other with 32 neurons. The 2 hidden layers had 25% dropout and the last layer had 50% dropout.

Recently, Kannadasan et al. [111] used a stacked autoencoder based DNN model to predict Diabetes. A typical autoencoder model consists of two components namely encoder and decoder. In this work, the authors used only the encoder part. In this model, the input of the autoencoder was the input of the encoder and the output of the encoder was the output of the hidden layer of the autoencoder. The DNN model was trained to minimize 3 errors: Mean Square Error (MSE), regularization and sparsity using backpropagation. Some researchers only tried to evaluate the performance of the dataset on a single classifier. For example, Rakshit et al. [173] implemented a 2-class DNN after standardization and missing value deletion, Daanouni et al. [46] compared four classifiers, namely DT, k-NN, ANN and DNN. In [46], DNN with FS using Neighbourhood Components Analysis (NCA) yielded the best result. On the contrary, a comparative analysis of different algorithms was performed by authors like Hashi et al. [84], and Sisodia and Sisodia [206]. Three shallow classifiers namely DT, SVM and NB were used by Sisodia and Sisodia [206], out of which NB yielded the best result.

7 CKD Prediction Methods

The methods considered in this survey use diagnostic data for CKD prediction based on the UCI CKD dataset[20]. The dataset has 25 diagnostic attributes and a single target variable. In all, there are 11 nominal and 13 categorical features. The research endeavors of the past about the UCI CKD dataset have been compared in Table 11. Some of the ideas presented in the existing works have been discussed here.

An important aspect of any classification system is the quality of data which can be enhanced by cleaning and preprocessing techniques, some of which include missing value imputation, outlier removal, and class imbalance rectification. These areas have been explored by a number of researchers in the case of designing diagnostic report based disease screening methods, which is true for CKD detection also. For example, Pujianto et al. [167] cleaned the data by creating 'pure clusters' using the k-means clustering algorithm. The clustering process generated two 'pure' clusters containing only those instances whose labels corresponded with the cluster label. Data not belonging to any of the pure clusters were removed prior to learning the classifier. The 'pure' cluster-merged dataset was validated using SVM with polynomial, RBF and Sigmoid kernels. This method is similar to the one used by Patil et al. [160] in Diabetes detection and thus had the same problem of dataset shrinkage as the former. In another work, Harimoorthy and Thangavelu [81] modified the redial basis kernel (discussed earlier) of SVM which in turn decreased the classification error of the SVM model.

The problem of class imbalance was scrutinized by Wibawa et al. [225]. They also introduced a novel method

---

[20] http://archive.ics.uci.edu/ml//datasets/Chronic_Kidney_Disease

of ELM using RBF, Linear, Polynomial and Wavelet kernels. Various combinations of imbalance correction, FS and kernels were applied. ELM with RBF kernel with undersampling correction FS yielded the best results. Though undersampling may reduce the imbalance in the dataset, it shrinks an already miniature dataset and sometimes makes the model prone to overfitting. Hence, using a hybrid of oversampling and undersampling algorithms may lead to better classification. To eradicate the problem of missing values, a three-fold approach was designed by Sisodia and Verma [207]. If the number of missing values was small for an attribute, then the respective instances were deleted. However, if the number of missing values was large in comparison to the total number of instances, the attribute itself was deleted. If the number of missing values was moderate, then they were imputed. Finally, the dataset was classified using NB, SVM, J48, and three ensemble classifiers, namely, RF, Bagging, and AdaBoost.

Aljaaf et al. [11] filtered the data to weed out outliers and also dealt with missing values. The outliers were replaced with the highest or lowest permissible values. For missing values, a two-fold approach was adopted. If the missing values comprised less than 20% of the total data, the data was deleted. Otherwise, the missing value was imputed. This was followed by FS which produced an optimal feature subset of 7 features. These were trained using Classification and Regression Trees (CARTs), LR, MLP and SVM classifiers. Another significant research direction is the predominance of FS algorithms owing primarily to a large number of features which leads to what can be called a twin curse: an increase in training time and the possibility of overfitting. A number of authors have tried different FS strategies like IG, Best First Search (BFS), etc. Polat et al. [164] demonstrated the utility of FS algorithms by comparing the performance of SVM with and without FS. By fixating on a single classification algorithm: SVM, the authors comprehensively illustrated that FS invariably improved the classification accuracy. The FS algorithms used were classifier subset evaluator, wrapper subset evaluator, correlation FS subset evaluator, and filtered subset evaluator which employs either best first search or greedy stepwise methodology. The filtered subset evaluator with the BFS method outperformed all other strategies.

Chetty et al. [41] used BFS for FS followed by classification. The reduced feature subset was used for CKD prediction using NB, SVM and k-NN classifiers, out of which k-NN produced the best result. Like Chetty et al. [41], Charleonnan et al. [35] used BFS algorithm for FS. The authors converted the nominal attributes to binary attributes in the training data followed by FS. k-NN, SVM, LR and DT were used as the classifiers out of which SVM yielded the best results. One of the disadvantages of the BFS is that it, being a greedy algorithm, may get stuck in local optima.

Hence, the selected feature subset may not be the optimal feature subset always. A popular method using the correlation between features and target variables has been used by Almansour et al. [12] and Wibawa et al. [226]. Almansour et al. [12] deleted the missing values. After the preprocessing, the best hyperparameters for ANN and SVM were selected. Using these optimized hyperparameters, the authors computed the correlations among all the features and the target which were then arranged in descending order. Half the features having the best correlation with the target variable were retained and tested with ANN and SVM. The process of halving and testing was continued until a single feature was left. The feature subset which exhibited the best performance was used to build the final model. It is noteworthy that the process of halving the number of features may lead to the selection of a non-optimum feature subset. Sandhiya and Palani [186] selected the best features using the ICRF-LCFS algorithm where those features are selected which contributed most in predicting good results. Before that, the authors employed the k-Means algorithm to group them. After that T-CNN is applied to this. Wibawa et al. [226] combined boosted classifier and FS for predicting CKD. The features were first selected using Correlation FS (CFS). For the boosted classifier, it used the base learners like NB, k-NN and SVM. Though collinearity is a very easy and readable way to judge the relevance of the feature, it may lead to multicollinearity as only the correlation of feature variables with the target variable was taken into account. This could result in the inclusion of redundant features.

Owing to superior prediction capabilities, boosting algorithms have become extremely popular of late. Works by Wibawa et al. [226], Sedighi et al. [190] and Islam and Ripon [18] which combined FS and boosted classifier bear testament to this fact. Sedighi et al. [190] undertook a comprehensive comparative study of different filter and wrapper based FS methods. The missing values were imputed using the k-NN model. Forward FS (FFS), Backward Feature Elimination (BFE), Bi-directional Search (BDS) and GA were used for FS and trained with AdaBoost using NB as the base learner. GA was found to perform the best among all the FS algorithms. However, according to some studies like Ting and Zheng [219], using a boosting algorithm on NB may not produce the desired improvement in classification compared to NB. Introducing a tree structure in NB followed by boosting can improve the classification significantly.

Apart from FS, feature extraction methods were also used to reduce dimensionality. Though not extensively explored in literature, authors like Guia et al. [74] have used PCA. In this work, after missing value imputation, label encoding and normalization, PCA was used for dimensionality reduction. The first 11 principal components accounting for more than 95% of the variance, were used. SVM, DT, RF, Gaussian NB, MLP, and k-NN were used as the classifiers out of

which k-NN yielded the best accuracy. It is noteworthy that PCA works best with strongly correlated features. Hence, using CFS followed by PCA could enhance the predictive power of the model. Some authors like Avci et al. [21] and Amirgaliyev et al. [15] emphasized the comparative study of various algorithms. Avci et al. [21] provided a simple comparison of various ML algorithms like NB, K-Star, SVM and J48. It was found that the J48 algorithm had the best performance. An SVM-based approach was used by Amirgaliyev et al. [15] where they used Linear SVM and the results were compared for 10-fold cross-validation. In another work, Johari et al. [108] made a comparative study of DT and 2-class NN after feature normalization using min-max normalization. The NN yielded much better results across various performance metrics. A similar study was done by Vashisth et al. [223] among MLP, SVM and NB where MLP outperformed the other algorithms.

8 Liver Disease Prediction Methods

The majority of the research works found in the literature for Liver Disease prognosis have used diagnosis report based classification utilizing the Indian Liver Patient Dataset (ILPD)[21] and Liver Disorders Dataset (LID)[22]. Both the datasets are available in the UCI repository. ILPD is a multivariate dataset consisting of 583 patient records (441 male and 142 female). All the patients were from the northeast part of Andhra Pradesh, India. It has 10 attributes and 416 out of 583 patients are classified as liver patients and the rest are non-liver patients in the dataset. LID (originally known as the British United Provident Association (BUPA) dataset) is a multivariate dataset consisting of 345 samples having 7 attributes. Each sample constitutes the record of a male subject. The subjects were labelled as either class label 1 (positive case) or 2 (negative case). Out of 345, there are 145 and 200 positive and negative subjects respectively.

There is a plethora of works found in the literature that have selected a classifier from a pool of available classifiers based on their merit for Liver disease screening while experimenting on IPLD and/or LTD. For example, Kant and Ansari [112] utilized the Atkinson index, a popular measure of inequality to select the initial centroid of the k-means algorithm. The improved k-means clustering with the Atkinson index produced a thriving precision score on the ILPD dataset. In another work, Adil et al. [4] used an LR classifier that is simple and possesses low computational cost. They split the ILPD into training and testing data in a 7:3 ratio and

applied k-fold cross-validation for partitioning the training data. Gogi and Vijayalakshmi [68] used various classifiers like SVM, LR and DT for the classifying samples belonging to ILPD. However, they applied missing value imputation with not a number (NAN) in the preprocessing stage. In their experiment, LR performed better than the rest and yielded the best classification metrics. In another work, Kumar and Katyal [120] reported that the C5.0 with adaptive boosting outperformed the classical classifiers like NB, RF, k-NN, actual C5.0 (a DT based classifier) and the k-means clustering algorithm in terms of accuracy, precision and recall values while detecting liver disease.

Contrary to the above-mentioned techniques, some researchers have tweaked the native classifiers to obtain better classification performance. For example, Tiwari et al. [220] employed various Lazy classifiers like instance based k-NN using Log and Gaussian (IBKLG) weight kernels, Locally Adaptive k-NN (Localk-NN) algorithms and Rough Set Exploration System Library based k-NN (Rseslibk-NN) for stratifying ILPD. Comparatively, Localk-NN showed the finest accuracy and recall measure while IBKLG overpowered the rest of the lazy classifiers in terms of precision score. However, it could not handle circumstances where the training data is insufficient or corrupted with noise. In another work, Chua and Tan [45] upgraded the classification capability K-NN classifier by introducing a fuzzy rule-based k-NN algorithm for Liver Disease classification utilizing the LID dataset. Unlike conventional k-NN, the proposed method's initialization procedure operated the imprecise inputs (neighborhood density and distance) through the natural framework of the fuzzy logic system. Euclidean distance measurement was adopted to attenuate dimensionality. Fuzzy rule-base's performance was enhanced by a binary coded GA that concomitantly optimizes the parameters of the rule consequent, antecedent membership functions and the feature weights. The decision boundary produced by the proposed method was tunable and thus it could handle circumstances where the training data is insufficient or corrupted with noise. After evaluation, it was evident that fuzzy rule-based k-NN outperformed crisp and fuzzy k-NN.

All the prior mentioned techniques followed a common trend of utilizing robust classifiers on the raw dataset without concluding the distinguishing capability of the attributes present in the dataset. However, some researchers have focused on the selection and ranking of the features and reported that the performance of most practical classifiers improves a lot when highly-correlated or irrelevant features are omitted from a given dataset. For example, Kulkarni and Shinde [118] came up with a hybrid neuro-fuzzy classification model to predict Liver Disease from the LID dataset where IG was used to select the feature subset that gives the highest classification accuracy. In this work, ANN was used to comprehend the membership value for fuzzy classes

---

[21] https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

[22] https://archive.ics.uci.edu/ml/datasets/liver+disorders

Table 12 Performance comparison of recent prediction models on UCI ILPD and UCI Liver Disorders datasets. Classifiers with '*' represent the classifiers with the highest ACC scores in the respective paper

Work Ref.	Classifier	Performance metrics					
		ACC	P	R	F1	S	AUC
Methods on UCI Indian Liver Patient dataset							
Adil et al. [4]	LR	0.74	0.72	0.74	0.69	–	–
Gogi and Vijayalakshmi [68]	LR*	0.95	–	–	–	–	0.93
Abdalrada et al. [1]	LR	0.72	–	0.90	–	0.78	0.75
Srivenkatesh [210]	LR*	0.76	0.77	0.95	0.85	–	–
Kant and Ansari [112]	k-means	0.23	0.94	0.31	0.46	–	–
Ramana and Boddu [174]	Bagging	0.69	0.44	0.95	–	0.85	0.70
Nahar and Ara [83]	SVM	0.73	–	0.96	–	0.96	–
Babu et al. [22]	NB*	0.56	0.76	–	–	0.95	–
Lakshmi et al. [124]	C4.5	–	–	0.78	–	–	–
Kumar and Katyal [120]	C5.0	0.75	0.90	0.78	–	–	–
Tiwari et al. [220]	k-NN	0.98	–	–	–	–	0.98
Kumar and Sahu [119]	RF*	0.79	0.69	0.61	0.65	0.87	–
Methods on UCI Liver Disorders dataset							
Chua and Tan [45]	k-NN	0.67	–	–	–	–	–
Kulkarni and Shinde [118]	ANN	0.67	–	–	–	–	–
Haque et al. [80]	ANN	0.85	–	0.85	0.82	0.89	–

of an input dataset. Next, they employed the sum aggregation reasoning rule to aggregate attribute belongingness so that the pattern belongingness to the given classes can be calculated. Obtained attributes were thereafter used for high level decision making first and then utilizing defuzzification operation, the pattern was assigned to the predicted class. Underlying attributes were ranked using some ranker method and then a few top ranked attributes which were relevant to making predictions were utilized to train the classification model.

Babu et al. [22] experimented with different attribute evaluator techniques available in the WEKA tool and applied them to ILPD while Nahar and Ara [83] used Singular Value Decomposition (SVD) based ranking algorithm available in Matrix Laboratory (MATLAB) for predicting Liver Disease. Going into detail, Babu et al. [22] employed a k-means clustering algorithm on the top-7 ranked features of the training dataset and checked the clustered validity of objects. Mis-clustered data were omitted from the training dataset. Finally, the refined train set was fed into various classifiers such as NB, K-NN and DT (C4.5 algorithm) for classification. NB classifier outdid k-NN and DT. However, Nahar and Ara [83] experimented with different top ranked features using SVM as a classifier. The authors experimentally showed that top-8 ranked features performed best for both the ILPD and the LID datasets. Pearson Correlation Coefficient (PCC) of the attributes was used by Haque et al. [80]. In their work, they used RF and ANN as a classifier in order to discern liver disorder patients from rest. Experimentally, NN outperformed RF. Abdalrada et al. [1] used a sigmoid function based feature ranking method. Using the sigmoid function occurrence probability (say, p) of each attribute was calculated. Attributes having a low p value (< 0.05) and odds ratio were considered to be less important and eventually they were removed. Only 5-attributes viz., Age, DB, SPGT, TP and Albumin were chosen for Liver Disease prediction using the LR classifier.

Another problem that plagues researchers is the presence of missing values. Choudhury and Pal [43] introduced a novel method of missing value imputation using autoencoder NN. The dataset was trained on multiple datasets like the ILPD without any missing values for the attribute. The trained autoencoder is then used to predict missing values. As an initial guess of the missing value, the nearest neighbor rule is used and then refined by minimizing the reconstruction error. This is based on the hypothesis that a good choice for a missing value would be the one that can reconstruct itself from the autoencoder. The classifiers used for the imputation are SVM, LR, NB, MLP, PRZ, k-NN, CART and PNN. The researchers comprehensively establish the superiority of their imputation strategy especially for high rates of missing values.

In the literature, some authors like Kumar and Sahu [119], Auxilia [20], Ramana and Boddu [174] and Srivenkatesh [210] emphasized the comparative performance of various classifiers on the ILPD. Kumar and Sahu [119] performed a set of experiments to check the performance of classifiers like RF, CART and Reduced Error Pruning Tree (REPTree) with two different train test splits: 70%-30% and 80%-20%, and inclusion/exclusion of FS technique. They obtained the

best result using the FS method with RF on 80%-20% on the train-test split. In [20], the author examined the performance of classifiers like DT, ANN and RF to foresee a liver ailment in patients. DT outperformed the rest of the classifiers based on some factual metrics. Srivenkatesh [210] investigated performances of five different supervised learning algorithms such as NB, LR, SVM, RF and k-NN. LR outperformed the rest of the classifiers. Performance measures like Kappa statistic value, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), P and F1-score were utilized to assess the performances of the classifiers. Also, a comparative study by Ramana and Boddu [174] includes the classifiers like Bagging, k-NN, DT (J48 algorithm), JRip (a RIPPER algorithm), MLP and NB classifiers. The bagging classifier achieved the highest accuracy followed by MLP, J48, JRip, k-NN and NB. These comparative studies allow us in perceiving the advantages and disadvantages of various classifiers and provide insight into various techniques.

Table 12 enumerates the performance of various research efforts over the last 10 years about Liver Disease prediction using ILPD and LID datasets.

9 Future Research Directions

From our thorough analysis of different computerized disease detection methods over the different datasets related to Breast Cancer, Lung Cancer, Leukemia, Heart Disease, kidney disorder (i.e, CKD) and Liver Disease, it is clear that AI and ML based approaches were extensively used by the researchers since last two decades. New methods are constantly being explored to develop better models for the prediction of diseases now and then. It comes as no surprise that the models designed in the last few years have comprehensively outperformed previous research attempts across all the datasets considered in our survey. While this survey does chronicle significant inroads of AI and ML in disease prediction and diagnosis, there are some significant research gaps that future researchers should take up. Some of them have been enlisted and described hereafter in this section.

9.1 Exploring more DL Models

Researchers primarily used ML (Pujianto et al. [167] in UCI CKD, Ilango and Ramaraj [99] in PIMA Indians, etc.) or Shallow NNs (Aljaaf et al. [11] in UCI CKD, Gavhane et al. [64] in UCI Heart dataset, etc.) for image based classification systems. It is noteworthy that most of these approaches employed ML or shallow NNs as classifiers. DNNs consisting of multiple hidden layers have not been explored much as compared to shallow learners. DNNs were mainly explored using transfer learning (work by Shafique et al. [196] in ALL_IDB dataset) for designing pathological

image based disease detection. This is probably due to the concise nature of the datasets and the bottleneck of computing capabilities. With the abundance of data at the disposal of researchers and the availability of superior computers, it is expected that DNNs will be explored exhaustively in the near future. Besides, fully trained DNNs like ResNet and Inception could be employed in the future. Also, some hybrid models (e.g., [65, 224]) that feed some handcrafted features to the DL model could be another future direction. In addition to these, ensembles of different DL models bearing some complementary information (e.g., [50, 51, 162]) could be applied to enhance the overall classification model.

9.2 Designing Sophisticated Data Processing Strategies

From the above discussion, it is evident that data preprocessing had played a vital role over the years to enhance the end performance of the detection models. Such preprocessing techniques included missing value imputation and feature dimension reduction through FS. In this subsection, we first revisit and analyse the existing works and then suggest some ways which might improve the existing ones.

9.2.1 Handling Missing Values

The existence of missing values is a very common problem in the case of diagnosis report based datasets as this is similar to survey data. All the related information might not be available for all subjects as medical practitioners may collect only information they consider relevant. However, in ML based systems, such missing values are detrimental during prediction as learning may be biased to the available data. To handle such scenarios, many researchers like Patil et al. [160] for PIMA Indians, Almansour et al. [12] for the UCI CKD dataset, and Choudhury and Pal [43] for ILPD opted for deleting the missing values prior to actual classification. In some cases, the missing values are substituted with zero [12] or the mean value of the feature for the rest of the data like Chen et al. [39] in PIMA Indians among others to handle such missing values. However, such crude techniques of missing value imputation or deletion may not be the right choice for some disease predictions. To perform missing value imputation, one might employ a clustering approach where data samples can be clustered without considering the attribute having missing values first and then the information from the same cluster data may be used to substitute the missing values. Additionally, some state-of-the-art methods as described in [126, 171, 191, 233] could be applied.

9.2.2 Exploring more Dimensionality Reduction Strategies

Many datasets have a large number of features but not all features are useful to generate the final output. Such redundant data only increases the computational complexity. Feature engineering is a process to generate a reduced dataset with new features which are produced from the existing raw dataset. However, we find that for data reduction, most researchers have focused on FS. Dimensionality reduction using feature engineering has not been extensively explored and is mostly restricted to only PCA like Ophir et al. [62] in Fred Hutchinson Cancer Research Center Dataset and Karayılan and Kılıç [113] in UCI Heart Disease or LDA as in Mishra et al. [144]. Other dimensionality techniques like isomap embedding [184], locally linear embedding [157], supervised locally linear embedding algorithm [232] among others could be useful.

9.2.3 Designing new FS Strategies

There is also a dearth of variety in optimization algorithms with most researchers using gradient descent. Very few researchers used GA in WBC dataset (Huang et al. [98]), ILPD dataset [45] and PIMA Indians datasets [221], AMO Algorithm in UCI CKD Dataset [18] or PSO in UCI Heart Disease [60], WBC dataset [234] etc. which yielded satisfactory results in respective research domains. It is imperative that researchers in future might extensively use one or more such optimization algorithms. To be specific, for small sized data, a filter method or ensemble of filter methods could be employed while for the large sized data, use of wrapper filter methods (i.e., filter method followed by wrapper methods) or hybrid wrapper methods or wrapper method with some local search could be another future scope for the researchers.

9.3 Fusing Handcrafted Features with Deep Features

For image based classification systems which typically need feature extraction followed by the classification, researchers mostly focused on shape, colour or texture based features. Some state-of-the-art feature extraction techniques like Scale Invariant Feature Transform, Speeded Up Robust Features, Oriented FAST and Rotated BRIEF (ORB), Daisy and Haralick (except [195, 222]) feature had not been explored much till now for such problems. Moreover, researchers had also overlooked the possibility of some intelligent hybrid models combining handcrafted features with deep features like [195]. Therefore, exploring the use of state-of-the-art handcrafted features or hybrid models fusing handcrafted features with deep features might help in improving the present state-of-the-art results obtained by image based CAD models.

9.4 Handling Class Imbalance Problem

Class imbalance is very common when considering medical data since the number of infected subjects is very low as compared to the number of healthy subjects. However, a very few researchers, as described during methodological description, have taken into account the problem of imbalanced class in the dataset (e.g., Han et al. [79] in BreakHis Dataset and Dutta et al. [54] in PIMA Indians) while using ML based approach. Imbalanced datasets can often lead to poor classification and misleading accuracy measures. Class imbalance could be rectified either using undersampling or oversampling or a hybrid of the two approaches in case of the diagnostic report based as well as image based classification problems. Besides, data augmentation for image based classification problems could be a better choice.

9.5 Generating Large-scale Datasets

It is observed from the methodological reviews of the existing techniques that there is a crisis of large-scale datasets of the various diseases. We find that the datasets used here are restricted to a certain age bracket, gender, or location. Such homogeneity in the dataset may lead to the problem of overfitting and the proposed model may not generalize well for real world scenarios. Hence, care must be taken to build datasets that include a heterogeneous sample space. This is also a pressing need in the world of the Internet of Things (IoT) where medical practitioners all over the world can work as a team to deal with complicated diseases. Some special categories of data could also be prepared for aged people with different co-morbidity or as in the current scenario COVID-19 patients having no prominent symptoms. Here, we describe some of the possible data requirements that could be prepared in the future.

9.5.1 New Data Collection

Breast Cancer: WBC dataset is the most widely used dataset in the literature on Breast Cancer detection. However, this dataset was last updated in 1991 which makes it outdated. The number of instances is quite low with only 699 instances. Moreover, some important attributes causing Breast Cancer like age, obesity, menopause, and family history of having Breast Cancer are also absent there. For image based classification, BACH and BreakHis are some of the prominent Breast Cancer histopathological image datasets widely used by the researchers. Various data augmentation techniques are used to cope with model overfitting due to data scarcity issues in these datasets. However, there are several problems associated with data augmentation like loss of discriminating features or information. The selection of a data augmentation approach should be performed

judiciously based on the dataset characteristics. Having more data is always the recommended panacea, as data augmentation also brings up the possibility of the elimination of some inherent properties from the image.

Lung Cancer: The Irvine dataset (last update: 1992) consists of only 32 instances and the MLCD dataset (last update: August 2002) contains a total of 96 instances where 86 primary lung adenocarcinomas samples and 10 non-neoplastic lung samples. Such a low sample count and age-old dataset motivated researchers to provide the LCD dataset (last update: 2017) having 1000 instances. This data is much better as compared to the other two while the number of samples is taken into account. However, this dataset is not enough in a practical scenario. Hence, more such dataset is required to be generated to improve the Lung Cancer detection from diagnosis report and thereby assist the medical experts to make proper decisions.

Leukemia: Golub dataset is the most common and one of the early gene expression based datasets and this was a frequently used dataset in Leukemia related research. It originated in 1999, which makes it quite old. Moreover, this dataset has only 'ALL' and 'AML' data while data from non-cancerous subjects are missing. Hence, preparing a new dataset containing gene expression of ALL, AML and non-cancerous subjects or adding gene expression data of non-cancerous subjects to this dataset could open up a new direction on Leukemia detection from gene expression data. There are only 72 instances which is a quite low number. The histopathological image dataset used for the classification of Leukemia is the ALL_IDB dataset which consists of histopathological images collected as long back as 2005. Hence the dataset has become outdated. With better microscopes and software available to capture the image under the slides, a newer dataset would lead to better models.

Heart: The Cleveland dataset is used in almost every heart related research work which was developed in 1988. This dataset has 303 instances and it is quite low from the ML point of view. This dataset contains some important attributes like smoking, Diabetes but ignores some potentially useful attributes. Moreover, this dataset was only collected from Cleveland city, USA (Ohio) and thus making the dataset highly localized. It would be wiser to create a dataset that includes subjects from countries like Turkmenistan, Kazakhstan, Kyrgyzstan, Mongolia, and Russia which are highly affected by it. So we need more globalized and updated data for more applicability of the devised method on that new dataset to real life.

Diabetes: The dataset used in Diabetes includes only females above 21 years with the maximum age being 65. All the subjects belonged to PIMA Indian heritage. Hence the sample space considered here is confined to narrow geography and of a particular gender. Hence such a dataset is outdated and does not lead to the generation of a robust model. A new dataset with subjects across locations, genders and age brackets is the need of the hour. Moreover, the dataset is too small with only 768 instances to generalize any method experimented on this dataset. Even within the existing dataset, the data is skewed with only 268 instances being positive. In addition to these, the dataset was last updated in 1990 which makes it obsolete in the current scenario.

CKD: The dataset used for CKD prediction is the UCI CKD dataset. The dataset recorded patients from Tamil Nadu in India over a period of 2 months. Hence, like in Diabetes, the dataset lacks diversity in terms of geography. A dataset more inclusive of heterogeneity would lead to generating better models for the prediction of CKD. The dataset has only 400 instances out of which 62% of the subjects had CKD. This points to the imbalance in the dataset. These issues create an urge towards the generation of a new and larger dataset covering the long period and large geographical locations.

Liver Disease: The LID (generated in 1990) and ILPD (generated in 2012) were widely used diagnostic datasets for Liver Disease detection using classical ML based approaches. Several challenges are associated with these datasets for clinical research. Both the datasets are quite outdated and demographically localized. The ILPD has both male and female subjects records while LID is limited to male subjects only. ILPD comprises 583 patient records and LID consists of scarcely 345 instances. DNNs are data-eating approaches that require bountiful amounts of training data. Hence, they failed to yield good results in ILPD and LID when compared to classical ML models.

9.5.2 Generating Synthetic Data

We have already mentioned that DL based models, in general, require a large amount of data for proper training of the model. We have seen that the availability of sufficient and useful data becomes a major concern for the researchers. In the case of medical data, things are even more cumbersome as collection or preparation of medical data is most of the time very costly and often time-consuming. Besides the assemblage, the data also needs to be annotated by medical experts. In 2014, Goodfellow et al. [71] introduced Generative Adversarial Networks (GANs). GAN is a generative model to generate new realistic data samples from the reference sample data that are not only similar to the examples but are indistinguishable as well. Over recent years, GAN has been gaining a lot of attention in the medical fraternity due to its impeccable features, namely, image synthesizing, de-noising, segmentation, reconstruction, data simulation and classification. GAN and its extensions can extricate the scarcity of labelled data problems by generating close to realistic data from existing samples. However, synthetic data generated if not nearly identical to real-world data, can affect

the quality of decision making. The only way to guarantee a GAN generator is generating accurate, realistic outputs is to test its performance on well-understood, human annotated validation data using the discriminator. Although synthetic data generation has become easier over time, real-world human annotated data remains a cornerstone of training data for CADDs.

9.5.3 Preparing Multi-modal Data

For diseases like Breast Cancer, Leukemia and Lung Cancer we observe that researchers, in the past, used either diagnostic reports or pathological images for screening these diseases in most of the cases. Unfortunately, to the best of our knowledge, no method has used both. This might be due to the non-availability of multi-modal patient data [128] in the public domain. By multi-modal data, we mean that both diagnosis reports and pathological images for each subject in the dataset. In this context, it is to be noted that recent IoT-enabled smart health care systems are suggesting the use of multi-modal data for better patient care and more accurate disease diagnosis. Moreover, Sánchez-Cauce et al. [185] showed that the use of patient information like age and sex with thermal images improves the overall performance of the breast cancer detection system they designed. However, the lack of such data in the public domain limits the evolution of computer aided multi-modal disease screening systems that can overcome challenges identified by Madabhushi et al. [128]. Hence, the preparation of publicly available multi-modal data can be another important future research direction.

9.6 Medical Image Segmentation

A medical image segmentation technique aims at distinguishing the pixels of an organ or a lesion from the background present in the pathological image. This is considered to be one of the most challenging tasks in medical image analysis, which is a major component of AI-based diagnosis models by researchers. A medical image segmentation based disease diagnosis system assists the experts by delivering critical information like the shape and volume of these organs [86]. Even automatic segmentation of organs/lesions can assist the physician to monitor their growth and make surgical decisions [26]. In the early days of medical image segmentation methods, researchers relied on traditional methods like edge detection, morphological operations, skeletonization and various statistical properties. However, in the latter stage, handcrafted features based classification and clustering dominated the domain for a long period. At this stage, designing and extracting better features remained the primary concern for developing such a method. In recent

years, researchers obtained considerable success in medical image segmentation using DL based approaches [86, 235]. Thus, in the past few years, DL aided medical image segmentation techniques have gained vast popularity in the research community. However, researchers of this domain encountered limited data both in terms of count and variety. To be specific, these data are scarcely annotated or weakly annotated [217] and this makes the medical image segmentation problem a GT-hard problem [193]. To overcome this issue, one may employ approaches like semi-supervised learning [229], contrast learning [107], domain adaptation that is similar to zero-shot learning [29], and Graph Neural Network (GNN) [179] aided learning for region and boundary aggregation [137] during the training process or can generate synthetic data for training using a simulation-supervision approach [192, 193].

9.7 Maintaining Ethical and Legal Aspects

AI in the medical domain started its journey in the early 1970s when researchers were fascinated in this domain because of its huge real-life applications [159]. Till then, the world has observed the development of several decision support systems keeping AI at the backbone. Even in many cases, it has been observed that the AI-assisted methods are outperforming human precision [150, 211]. However, the performance of these methods is largely dependent on the training data and thus prejudiced samples in the training set might hamper the end outcome. Hence, data verification and validation by some medical experts while preparing such training samples are needed from ethical perspectives. Even the consent of subjects, who are undergoing such experiments, must be taken before conducting experiments and releasing their data in public. Moreover, a notable loophole in the existing AI-assisted methods is that the majority of the algorithms fail to explain the reason behind their prediction to the medical practitioners. Therefore, researchers of a newly developed AI-assisted algorithm in the medical domain should strongly make their algorithm explainable on ethical grounds. Such initiatives would help the medical experts to spot any error present therein, and thereby would increase the reliability of the algorithm. Furthermore, developed systems must mention which parts of a decision or what kinds of actions are taken by the AI based system. Any hypothesis and consideration made by the AI-assisted algorithm must be supported by some state-of-the-art medical findings. The communication required for a designed system needs to be transparent i.e., the precautions that should be taken and measurable damage due to malfunction are to be mentioned clearly. Robustness of such systems is also expected. Interested readers are suggested reading the article

by Muller et al. [150] to know more about ethical aspects related to practical AI assisted medical systems.

The legal aspect is another big concern when developing and deploying any AI assisted systems. The present regulations and acts (e.g., Artificial Intelligence Act[23]) related to legal consequences are mainly designed to protect fundamental rights, also known as human rights, as an erroneous outcome from an AI-assisted medical system can cause serious physical and mental consequences of a subject. However, the present regulations and acts may be insufficient considering the evolving nature of AI in the medical system. Even in many cases, for example, the Artificial Intelligence Act, the accountability and liability for the damage caused by an AI-assisted medical system is not addressed clearly [211]. Hence, in the future, more insightful discussions on the current legal aspects such as regulations, acts and recommendations can help increase the reliability and end users' acceptability of the AI based medical systems.

10 Conclusion

Chronic diseases generally persist for a longer period in human organs and most importantly, there are no vaccines to prevent these diseases. However, early detection of these diseases can save many human lives. Nowadays, technology plays a significant role in healthcare systems, and there are many computer-aided supporting systems available. Owing to the huge social impact of this research field, many researchers from various domains like computer science, biology, medicine and statistics have contributed to this domain. To this end, the present survey makes a significant effort to record the current trends in research on chronic diseases like Breast Cancer, Leukemia, Lung Cancer, Heart Diseases, Diabetes, CKD and Liver. It covers methods including missing value imputation, feature reduction, FS, classifier combination, and ML and DL based approaches. The present survey puts more efforts on automatic disease prediction systems that used ML and DL based methods for the prediction of the said diseases. Moreover, this survey discusses the methods that considered diagnostic report based and images based inputs. Along with this, it mentions the advantages and limitations of the methods reported in the literature wherever possible. Additionally, this paper presents various comparative studies of the methods proposed by different researchers. Though some significant developments have been made in the automation of AI-enabled decision support systems, there are still some prominent gaps that hinder the systems from using them for practical purposes. Hence, this survey also tries to bring out some future

directions that need to be considered by researchers in order to make the systems usable for medical professionals.

Declarations

Conflict of interest All the authors declare that they have no conflict of interest or competing interests

References

1. Abdalrada AS, Yahya OH, Alaidi AHM, Hussein NA, Alrikabi HT, Al-Quraishi TAQ (2019) A predictive model for liver disease progression based on logistic regression algorithm. Periodicals of Engineering and Natural Sciences (PEN) 7(3):1255, https://doi.org/10.21533/pen.v7i3.667,

2. Abdeldaim AM, Sahlol AT, Elhoseny M, Hassanien AE (2017) Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis. In: Advances in Soft Computing and Machine Learning in Image Processing, Springer International Publishing, pp 131–147. https://doi.org/10.1007/978-3-319-63754-9_7,

3. Abdul-Aziz AA, Desikan P, Prabhakaran D, Schroeder LF (2019) Tackling the burden of cardiovascular diseases in India. Circulation. https://doi.org/10.1161/circoutcomes.118.005195

4. Adil SH, Ebrahim M, Raza K, Ali SSA, Hashmani MA (2018) Liver patient classification using logistic regression. In: 2018 4th International Conference on Computer and Information Sciences (ICCOINS), IEEE, pp 1–5. https://doi.org/10.1109/iccoins.2018.8510581,

5. Ahmadi H, Gholamzadeh M, Shahmoradi L, Nilashi M, Rashvand P (2018) Diseases diagnosis using fuzzy logic methods: a systematic and meta-analysis review. Comput Methods Programs Biomed 161:145–172

6. Ahmed S, Ghosh KK, Singh PK, Geem ZW, Sarkar R (2020) Hybrid of Harmony Search Algorithm and Ring Theory-Based Evolutionary Algorithm for Feature Selection. IEEE Access 8:102629–102645

7. Akter O, Moni MA, Islam MM, Quinn JM, Kamal A (2021) Lung cancer detection using enhanced segmentation accuracy. Appl Intell 51(6):3391–3404

8. Al-Milli N (2013) Backpropagation neural network for prediction of heart disease. J Theor Appl Inf Technol 56(1):131–135

9. Al-Tarawneh MS (2012) Lung cancer detection using image processing techniques. Leonardo Electron J Pract Technol 11(21):147–58

10. Alam J, Alam S, Hossan A (2018) Multi-stage lung cancer detection and prediction using multi-class SVM classifie. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, pp 1–4. https://doi.org/10.1109/ic4me2.2018.8465593,

11. Aljaaf AJ, Al-Jumeily D, Haglan HM, Alloghani M, Baker T, Hussain AJ, Mustafina J (2018) Early prediction of chronic kidney disease using machine learning supported by predictive

---

23 https://bit.ly/3AcDsCa

analytics. In: 2018 IEEE Congress on Evolutionary Computation (CEC), IEEE, pp 1–9. https://doi.org/10.1109/cec.2018.8477876,

12. Almansour NA, Syed HF, Khayat NR, Altheeb RK, Juri RE, Alhiyafi J, Alrashed S, Olatunji SO (2019) Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study. Comput Biol Med 109:101–111. https://doi.org/10.1016/j.compbiomed.2019.04.017

13. Alzubaidi MA, Otoom M, Jaradat H (2021) Comprehensive and comparative global and local feature extraction framework for lung cancer detection using ct scan images. IEEE Access 9:158140–158154

14. Amin MS, Chiam YK, Varathan KD (2019) Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform 36:82–93. https://doi.org/10.1016/j.tele.2018.11.007

15. Amirgaliyev Y, Shamiluulu S, Serek A (2018) Analysis of chronic kidney disease dataset by applying machine learning methods. In: 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), IEEE, pp 1–4. https://doi.org/10.1109/icaict.2018.8747140,

16. Anifah L, Haryanto, Harimurti R, Permatasari Z, Rusimamto PW, Muhamad AR (2017) Cancer lungs detection on CT scan image using artificial neural network backpropagation based gray level coocurrence matrices feature. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, pp 327–332. https://doi.org/10.1109/icacsis.2017.8355054,

17. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, Marami B, Prastawa M, Chan M, Donovan M et al (2019) Bach: grand challenge on breast cancer histology images. Med Image Anal 56:122–139

18. Arif-Ul-Islam, Ripon SH (2019) Rule induction and prediction of chronic kidney disease using boosting classifiers, ant-miner and j48 decision tree. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp 1–6. https://doi.org/10.1109/ecace.2019.8679388,

19. Ashiquzzaman A, Tushar AK, Islam MR, Shon D, Im K, Park JH, Lim DS, Kim J (2017) Reduction of overfitting in diabetes prediction using deep learning neural network. In: IT Convergence and Security 2017, Springer Singapore, pp 35–43. https://doi.org/10.1007/978-981-10-6451-7_5

20. Auxilia LA (2018) Accuracy prediction using machine learning techniques for indian patient liver disease. In: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, pp 45–50. https://doi.org/10.1109/icoei.2018.8553682

21. Avci E, Karakus S, Ozmen O, Avci D (2018) Performance comparison of some classifiers on chronic kidney disease data. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS), IEEE, pp 1–4. https://doi.org/10.1109/isdfs.2018.8355392

22. Babu MP, Ramjee M, Katta S, K S (2016) Implementation of partitional clustering on ILPD dataset to predict liver disorders. In: 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, pp 1094–1097. https://doi.org/10.1109/icsess.2016.7883256

23. Bache K, Lichman M (2013) Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of California. School of Information and Computer Science 28

24. Baldomero-Naranjo M, Martinez-Merino LI, Rodriguez-Chia AM (2021) A robust svm-based approach with feature selection and outliers detection for classification problems. Expert Syst Appl 178:115017

25. Baliarsingh SK, Muhammad K, Bakshi S (2021) SARA: a memetic algorithm for high-dimensional biomedical data. Appl Soft Comput 101:107009

26. Basak H, Kundu R, Sarkar R (2022) MFSNet: a multi focus segmentation network for skin lesion segmentation. Pattern Recognition, 108673

27. Begum S, Chakraborty D, Sarkar R (2016) Identifying cancer biomarkers from leukemia data using feature selection and supervised learning. In: 2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI), IEEE, pp 249–253. https://doi.org/10.1109/CMI.2016.7413749

28. Bhowal P, Sen S, Velasquez JD, Sarkar R (2022) Fuzzy ensemble of deep learning models using choquet fuzzy integral, coalition game and information theory for breast cancer histology classification. Expert Syst Appl 190:116167

29. Bian C, Yuan C, Ma K, Yu S, Wei D, Zheng Y (2021) Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation. IEEE Trans Med Imaging 41(5):1043–1056

30. Brancati N, Frucci M, Riccio D (2018) Multi-classification of breast cancer histology images by using a fine-tuning strategy. In: International conference image analysis and recognition. Springer, pp 771–778. https://doi.org/10.1007/978-3-319-93000-8_87

31. Brooks JP (2011) Support vector machines with the ramp loss and the hard margin loss. Oper Res 59(2):467–479

32. Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F, Manuel DG (2020) A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med Inform Decis Mak 20(1):1–12

33. Chakraborty B, Chaterjee A, Malakar S, Sarkar R (2022) An iterative approach to unsupervised outlier detection using ensemble method and distance-based data filterings. Complex Intell Syst. https://doi.org/10.1007/s40747-022-00674-0

34. Chakraborty B, Chaterjee A, Malakar S, Sarkar R (2022) An iterative approach to unsupervised outlier detection using ensemble method and distance-based data filtering. Complex Intell Syst. https://doi.org/10.1007/s40747-022-00674-0

35. Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N (2016) Predictive analytics for chronic kidney disease using machine learning techniques. In: 2016 Management and Innovation Technology International Conference (MITicon), IEEE, pp MIT–80. https://doi.org/10.1109/miticon.2016.8025242

36. Chatterjee B, Bhattacharyya T, Ghosh KK, Singh PK, Geem ZW, Sarkar R (2020) Late acceptance hill climbing based social ski driver algorithm for feature selection. IEEE Access 8:75393–75408

37. Chatterjee S, Biswas S, Majee A, Sen S, Oliva D, Sarkar R (2021) Breast cancer detection from thermal images using a grunwald-letnikov-aided dragonfly algorithm-based deep feature selection method. Comput Biol Med 141:105027

38. Chattopadhyay S, Dey A, Singh PK, Sarkar R (2022) Drda-net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images. Comput Biol Med 145:105437

39. Chen W, Chen S, Zhang H, Wu T (2017) A hybrid prediction model for type 2 diabetes using k-means and decision tree. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, pp 386–390. https://doi.org/10.1109/icsess.2017.8342938

40. Chennamsetty SS, Safwan M, Alex V (2018) Classification of breast cancer histology image using ensemble of pre-trained neural networks. In: International Conference Image Analysis and Recognition, Springer, pp 804–811. https://doi.org/10.1007/978-3-319-93000-8_91

41. Chetty N, Vaisla KS, Sudarsan SD (2015) Role of attributes selection in classification of chronic kidney disease patients. In:

2015 International Conference on Computing, Communication and Security (ICCCS), IEEE, pp 1–6. https://doi.org/10.1109/cccs.2015.7374193

42. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258

43. Choudhury SJ, Pal NR (2019) Imputation of missing data with neural networks for classification. Knowl-Based Syst. https://doi.org/10.1016/j.knosys.2019.07.009

44. Chouhan N, Khan A, Shah JZ, Hussnain M, Khan MW (2021) Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography. Comput Biol Med 132:104318

45. Chua T, Tan W (2009) A new fuzzy rule-based initialization method for k-nearest neighbor classifier. In: 2009 IEEE International Conference on Fuzzy Systems, IEEE, pp 415–420. https://doi.org/10.1109/fuzzy.2009.5277215

46. Daanouni O, Cherradi B, Tmiri A (2020) Diabetes diseases prediction using supervised machine learning and neighbourhood components analysis. In: Proceedings of the 3rd International Conference on Networking, Information Systems & Security, ACM, pp 1–5. https://doi.org/10.1145/3386723.3387887

47. Danjuma KJ (2015) Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. arXiv preprint arXiv:1504.04646

48. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 36(4):7675–7680. https://doi.org/10.1016/j.eswa.2008.09.013

49. Dey C, Bose R, Ghosh KK, Malakar S, Sarkar R (2021) Lagoa: learning automata based grasshopper optimization algorithm for feature selection in disease datasets. J Ambient Intell Hum Comput 13(6):3175–3194

50. Dey S, Bhattacharya R, Malakar S, Mirjalili S, Sarkar R (2021) Choquet fuzzy integral-based classifier ensemble technique for covid-19 detection. Comput Biology Med 135:104585

51. Dey S, Roychoudhury R, Malakar S, Sarkar R (2022) An optimized fuzzy ensemble of convolutional neural networks for detecting tuberculosis from chest x-ray images. Appl Soft Comput 114:108094

52. Dey S, Roychoudhury R, Malakar S, Sarkar R (2022b) Screening of breast cancer from thermogram images by edge detection aided deep transfer learning model. Multimedia Tools and Applications pp 1–19

53. Doppalapudi S, Qiu RG, Badr Y (2021) Lung cancer survival period prediction and understanding: deep learning approaches. Int J Med Inform 148:104371

54. Dutta D, Paul D, Ghosh P (2018) Analysing feature importances for diabetes prediction using machine learning. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, pp 924–928, https://doi.org/10.1109/iemcon.2018.8614871

55. Dwivedi AK (2018) Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Comput Appl 29(12):1545–1554. https://doi.org/10.1007/s00521-016-2701-1

56. Dwivedi AK (2018) Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl 29(10):685–693. https://doi.org/10.1007/s00521-016-2604-1

57. Fatima M, Pasha M et al (2017) Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl 9(01):1

58. Feig SA, Yaffe MJ (1995) Digital mammography, computer-aided diagnosis, and telemammography. Radiol Clin N Am 33(6):1205

59. Fernández-Alemán JL, Carrillo-de Gea JM, Hosni M, Idri A, García-Mateos G (2019) Homogeneous and heterogeneous ensemble classification methods in diabetes disease: a review. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp 3956–3959

60. Feshki MG, Shijani OS (2016) Improving the heart disease diagnosis by evolutionary algorithm of pso and feed forward neural network. In: 2016 Artificial Intelligence and Robotics (IRANOPEN), IEEE, pp 48–53. https://doi.org/10.1109/RIOS.2016.7529489

61. Fu L et al (2015) The discriminate analysis and dimension reduction methods of high dimension. Open J Soc Sci 3(03):7. https://doi.org/10.4236/jss.2015.33002

62. Gal O, Auslander N, Fan Y, Meerzaman D (2019) Predicting complete remission of acute myeloid leukemia: machine learning applied to gene expression. Cancer Inform 18:1176935119835544. https://doi.org/10.1177/1176935119835544

63. Gao X, Liu X (2018) A novel effective diagnosis model based on optimized least squares support machine for gene microarray. Appl Soft Comput 66:50–59. https://doi.org/10.1016/j.asoc.2018.02.009

64. Gavhane A, Kokkula G, Pandya I, Devadkar K (2018) Prediction of heart disease using machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, pp 1275–1278. https://doi.org/10.1109/ICECA.2018.8474922

65. Georgescu MI, Ionescu RT, Popescu M (2019) Local learning with deep and handcrafted features for facial expression recognition. IEEE Access 7:64827–64836

66. Ghosh KK, Singh PK, Hong J, Geem ZW, Sarkar R (2020) Binary social mimic optimization algorithm with x-shaped transfer function for feature selection. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2996611

67. Ghosh KK, Guha R, Bera SK, Kumar N, Sarkar R (2021) S-shaped versus V-shaped transfer functions for binary Manta ray foraging optimization in feature selection problem. Neural Comput Appl 33(17):11027–11041

68. Gogi VJ, MN V (2018) Prognosis of liver disease: using machine learning algorithms. In: 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), IEEE, pp 875–879. https://doi.org/10.1109/icrieece44171.2018.9008482

69. Gokulnath CB, Shantharajah S (2019) An optimized feature selection based on genetic approach and support vector machine for heart disease. Clust Comput 22(6):14777–14787. https://doi.org/10.1007/s10586-018-2416-4

70. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537. https://doi.org/10.1126/science.286.5439.531

71. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. 1406.2661

72. Guha R, Ghosh M, Chakrabarti A, Sarkar R, Mirjalili S (2020) Introducing clustering based population in Binary Gravitational Search Algorithm for Feature Selection. Appl Soft Comput 93:106341

73. Guha R, Ghosh M, Mutsuddi S, Sarkar R, Mirjalili S (2020) Embedded chaotic whale survival algorithm for filter-wrapper feature selection. Soft Comput 24(17):12821–12843

74. de Guia JD, Concepcion RS, Bandala AA, Dadios EP (2019) Performance comparison of classification algorithms for diagnosing chronic kidney disease. In: 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology,

Communication and Control, Environment, and Management (HNICEM), IEEE, pp 1–7. https://doi.org/10.1109/hnicem48295. 2019.9073568

75. Gultepe Y (2021) Performance of Lung Cancer Prediction Methods Using Different Classification Algorithms. CMC-COMPUTERS MATERIALS & CONTINUA 67(2)

76. Gunarathne W, Perera K, Kahandawaarachchi K (2017) Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, pp 291–296. https://doi.org/10.1109/bibe.2017.00-39

77. Gupta K, Chawla N (2020) Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained CNN. Procedia Comput Sci 167:878–889. https://doi.org/10.1016/j.procs.2020.03.427

78. Gupta S, Gupta MK (2021) A comparative analysis of deep learning approaches for predicting breast cancer survivability. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-021-09679-317

79. Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S (2017) Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 7(1):1–10

80. Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK (2018) Performance evaluation of random forests and artificial neural networks for the classification of liver disorder. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, pp 1–5. https://doi.org/10.1109/ic4me2.2018.8465658

81. Harimoorthy K, Thangavelu M (2021) Multi-disease prediction model using improved svm-radial bias technique in healthcare monitoring system. J Ambient Intell Hum Comput 12(3):3715–3723

82. Hasan MK, Alam MA, Das D, Hossain E, Hasan M (2020) Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8:76516–76531. https://doi.org/10.1109/access.2020.2989857

83. Hashem EM, Mabrouk MS (2014) A study of support vector machine algorithm for liver disease diagnosis. Am J Intell Syst 4:9–14

84. Hashi EK, Zaman MSU, Hasan MR (2017) An expert clinical decision support system to predict disease using classification techniques. In: 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp 396–400. https://doi.org/10.1109/ecace.2017.7912937

85. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

86. Hesamian MH, Jia W, He X, Kennedy P (2019) Deep learning techniques for medical image segmentation: achievements and challenges. J Digit Imaging 32(4):582–596

87. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507. https://doi.org/10.1126/science.1127647

88. Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11(1):63–90. https://doi.org/10.1023/A:1022631118932

89. Holzinger A (2016) Machine Learning for Health Informatics. In: Machine learning for health informatics, Springer, pp 1–24

90. Hosni M, Abnane I, Idri A, de Gea JMC, Alemán JLF (2019) Reviewing ensemble classification methods in breast cancer. Comput Methods Programs Biomed 177:89–112

91. Hosni M, Carrillo-de Gea JM, Idri A, Fernández-Alemán JL, García-Berná JA (2019b) Using ensemble classification methods in lung cancer disease. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp 1367–1370

92. Hosni M, García-Mateos G, Carrillo-de Gea JM, Idri A, Fernández-Alemán JL (2020) A mapping study of ensemble classification methods in lung cancer decision support systems. Med Biol Eng Comput 58(10):2177–2193

93. Houssein EH, Emam MM, Ali AA, Suganthan PN (2021) Deep and machine learning techniques for medical imaging-based breast cancer: a comprehensive review. Expert Syst Appl 167:114161

94. Hsieh SH, Wang Z, Cheng PH, Lee IS, Hsieh SL, Lai F (2010) Leukemia cancer classification based on support vector machine. In: 2010 8th IEEE International Conference on Industrial Informatics, IEEE, pp 819–824. https://doi.org/10.1109/INDIN.2010.5549638

95. Hsu CH, Chen X, Lin W, Jiang C, Zhang Y, Hao Z, Chung YC (2021) Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning. Measurement 175:109145

96. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017a) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4700–4708

97. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501. https://doi.org/10.1016/j.neucom.2005.12.126

98. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF (2017) Svm and svm ensembles in breast cancer prediction. PLoS ONE 12(1):e0161501. https://doi.org/10.1371/journal.pone.0161501

99. Ilango BS, Ramaraj N (2010) A hybrid prediction model with f-score feature selection for type II diabetes databases. In: Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India—A2CWiC '10, ACM Press, pp 1–4. https://doi.org/10.1145/1858378.1858391

100. Ismaeel S, Miri A, Chourishi D (2015) Using the extreme learning machine (elm) technique for heart disease diagnosis. In: 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), IEEE, pp 1–3

101. Jabbar M, Deekshatulu B, Chandra P (2015) Computational intelligence technique for early diagnosis of heart disease. In: 2015 IEEE International Conference on Engineering and Technology (ICETECH), IEEE, pp 1–6. https://doi.org/10.1109/ICETECH.2015.7275001,

102. Jahangir M, Afzal H, Ahmed M, Khurshid K, Nawaz R (2017) An expert system for diabetes prediction using auto tuned multi-layer perceptron. In: 2017 Intelligent Systems Conference (IntelliSys), IEEE, pp 722–728. https://doi.org/10.1109/intellisys.2017.8324209,

103. Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: a review. Egypt Inform J 19(3):179–189

104. Jakimovski G, Davcev D (2018) Lung cancer medical image recognition using deep neural networks. In: 2018 Thirteenth International Conference on Digital Information Management (ICDIM), IEEE, pp 1–5. https://doi.org/10.1109/icdim.2018.8847136

105. Jannesari M, Habibzadeh M, Aboulkheyr H, Khosravi P, Elemento O, Totonchi M, Hajirasouliha I (2018) Breast cancer histopathological image classification: a deep learning approach. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, p 242818. https://doi.org/10.1109/bibm.2018.8621307

106. Jayaraman V, Sultana HP (2019) Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-019-01193-6

107. Jog A, Carass A, Roy S, Pham DL, Prince JL (2015) MR image synthesis by contrast learning on neighborhood ensembles. Med Image Anal 24(1):63–76

108. Johari AA, Wahab MHA, Mustapha A (2019) Two-class classification: Comparative experiments for chronic kidney disease. In: 2019 4th International Conference on Information Systems and Computer Networks (ISCON), IEEE, pp 789–792. https://doi.org/10.1109/iscon47742.2019.9036306

109. Kamel SR, YaghoubZadeh R, Kheirabadi M (2019) Improving the performance of support-vector machine by selecting the best features by gray wolf algorithm to increase the accuracy of diagnosis of breast cancer. J Big Data 6(1):90. https://doi.org/10.1186/s40537-019-0247-7

110. Kanchan BD, Kishor MM (2016) Study of machine learning algorithms for special disease prediction using principal of component analysis. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), IEEE, pp 5–10. https://doi.org/10.1109/ICGTSPICC.2016.7955260

111. Kannadasan K, Edla DR, Kuppili V (2019) Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clin Epidemiol Global Health 7(4):530–535. https://doi.org/10.1016/j.cegh.2018.12.004

112. Kant S, Ansari IA (2015) An improved k means clustering with Atkinson index to classify liver patient dataset. Int J Syst Assurance Eng Manag 7(S1):222–228. https://doi.org/10.1007/s13198-015-0365-3

113. Karayılan T, Kılıç Ö (2017) Prediction of heart disease using neural network. In: 2017 International Conference on Computer Science and Engineering (UBMK), IEEE, pp 719–723. https://doi.org/10.1109/UBMK.2017.8093512

114. Kavitha K, Gopinath A, Gopi M (2017) Applying improved svm classifier for leukemia cancer classification using fcbf. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, pp 61–66. https://doi.org/10.1109/ICACCI.2017.8125817

115. Khandekar R, Shastry P, Jaishankar S, Faust O, Sampathila N (2021) Automated blast cell detection for Acute Lymphoblastic Leukemia diagnosis. Biomed Signal Process Control 68:102690

116. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680. https://doi.org/10.1126/science.220.4598.671

117. Kulkarni A, Panditrao A (2014) Classification of lung cancer stages on CT scan images using image processing. In: 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, IEEE, pp 1384–1388. https://doi.org/10.1109/icaccct.2014.7019327

118. Kulkarni UV, Shinde SV (2013) Hybrid fuzzy classifier based on feature-wise membership given by artificial neural network. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, pp 1–7. https://doi.org/10.1109/icccnt.2013.6726549

119. Kumar A, Sahu N (2017) Categorization of liver disease using classification techniques. Int J Res Appl Sci Eng Technol 5:826–828

120. Kumar S, Katyal S (2018) Effective analysis and diagnosis of liver disorder by data mining. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, pp 1047–1051. https://doi.org/10.1109/icirca.2018.8596817

121. Kumar V, Mishra BK, Mazzara M, Thanh DN, Verma A (2020) Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In: Advances in Data Science and Management, Springer, pp 435–442. https://doi.org/10.1007/978-981-15-0978-0_43

122. Kumari M, Singh V (2018) Breast cancer prediction system. Procedia Comput Sci 132:371–376. https://doi.org/10.1016/j.procs.2018.05.197

123. Labbé M, Martínez-Merino LI, Rodríguez-Chía AM (2019) Mixed integer linear programming for feature selection in support vector machine. Discret Appl Math 261:276–304

124. Lakshmi J, Jayanthi K, Sathya M (2017) Investigating ilpd for most significant features. Int J Mech Eng Technol 8:741–749

125. Li D, Zhang G, Wu Z, Yi L (2010) An edge embedded marker-based watershed algorithm for high spatial resolution remote sensing image segmentation. IEEE Trans Image Process 19(10):2781–2787. https://doi.org/10.1109/tip.2010.2049528

126. Lin WC, Tsai CF (2020) Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 53(2):1487–1509

127. Lv C, Chen C, Chen Q, Zhai H, Zhao L, Guo Y, Wang N (2019) Multiple pregnancies and the risk of diabetes mellitus in postmenopausal women. Menopause 26(9):1010–1015

128. Madabhushi A, Agner S, Basavanhally A, Doyle S, Lee G (2011) Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. Comput Med Imaging Graph 35(7–8):506–514

129. Mafarja MM, Mirjalili S (2017) Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing 260:302–312. https://doi.org/10.1016/j.neucom.2017.04.053

130. Maji S, Arora S (2019) Decision tree algorithms for prediction of heart disease. In: Information and Communication Technology for Competitive Strategies, Springer, pp 447–454. https://doi.org/10.1007/978-981-13-0586-3_45

131. Malakar S, Sharma P, Singh PK, Das M, Sarkar R, Nasipuri M (2017) A holistic approach for handwritten Hindi word recognition. Int J Comput Vis Image Process 7(1):59–78

132. Maleki N, Zeinali Y, Niaki STA (2021) A k-nn method for lung cancer prognosis with the use of a genetic algorithm for feature selection. Expert Syst Appl 164:113981

133. Mandal M, Singh PK, Ijaz MF, Shafi J, Sarkar R (2021) A tri-stage wrapper-filter feature selection framework for disease classification. Sensors 21(16):5571

134. Manhas J, Gupta RK, Roy PP (2021) A review on automated cancer detection in medical images using machine learning and deep learning based computational techniques: Challenges and opportunities. Arch Comput Methods Eng 1:1–41

135. Mansourypoor F, Asadi S (2017) Development of a reinforcement learning-based evolutionary fuzzy rule-based system for diabetes diagnosis. Comput Biol Med 91:337–352. https://doi.org/10.1016/j.compbiomed.2017.10.024

136. Medhekar DS, Bote MP, Deshmukh SD (2013) Heart disease prediction system using naive bayes. Int J Enhanced Res Sci Technol Eng 2(3)

137. Meng Y, Zhang H, Zhao Y, Yang X, Qiao Y, MacCormick IJ, Huang X, Zheng Y (2021) Graph-based region and boundary aggregation for biomedical image segmentation. IEEE Trans Med Imaging 41(3):69–701

138. Miao KH, Miao JH, Miao GJ (2016) Diagnosing coronary heart disease using ensemble machine learning. International Journal of Advanced Computer Science and Applications 7(10):30–39, https://doi.org/10.14569/IJACSA.2016.071004

139. Mirjalili S (2015) The ant lion optimizer. Adv Eng Softw 83:80–98. https://doi.org/10.1016/j.advengsoft.2015.01.010

140. Mirjalili S, Lewis A (2016) The whale optimization algorithm. Adv Eng Softw 95:51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008

141. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007

142. Mishra S, Sharma L, Majhi B, Sa PK (2016) Microscopic image classification using DCT for the detection of acute lymphoblastic leukemia (ALL). In: Advances in Intelligent Systems and Computing, Springer Singapore, pp 171–180. https://doi.org/10.1007/978-981-10-2104-6_16

143. Mishra S, Majhi B, Sa PK, Sharma L (2017) Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection. Biomed Signal Process Control 33:272–280. https://doi.org/10.1016/j.bspc.2016.11.021

144. Mishra S, Majhi B, Sa PK (2019) Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection. Biomed Signal Process Control 47:303–311. https://doi.org/10.1016/j.bspc.2018.08.012

145. Mohammed ZF, Abdulla AA (2021) An efficient cad system for all cell identification from microscopic blood images. Multimed Tools Appl 80(4):6355–6368

146. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7:81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

147. Mondal R, Malakar S, Barney Smith EH, Sarkar R (2022) Handwritten English word recognition using a deep learning based object detection architecture. Multimed Tools Appl 81(1):975–1000

148. Moon AM, Singal AG, Tapper EB (2019) Contemporary epidemiology of chronic liver disease and cirrhosis. Clin Gastroenterol Hepatol. https://doi.org/10.1016/j.cgh.2019.07.060

149. Moorthy U, Gandhi UD (2020) A novel optimal feature selection technique for medical data classification using anova based whale optimization. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02592-w

150. Muller H, Mayrhofer MT, Van Veen EB, Holzinger A (2021) The ten commandments of ethical medical AI. Computer 54(07):119–123

151. Murty NR, Babu MP (2017) A critical study of classification algorithms for lungcancer disease detection and diagnosis. Int J Comput Intell Res 13(5):1041–1048

152. Nadkarni NS, Borkar S (2019) Detection of lung cancer in CT images using image processing. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, pp 863–866. https://doi.org/10.1109/icoei.2019.8862577

153. Nasser IM, Abu-Naser SS (2019) Lung cancer detection using artificial neural network. Int J Eng Inf Syst 3(3):17–23

154. NirmalaDevi M, Appavu S, Swathi UV (2013) An amalgam KNN to predict diabetes mellitus. In: 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), IEEE, pp 691–695. https://doi.org/10.1109/ice-ccn.2013.6528591

155. Ogurtsova K, da Rocha Fernandes J, Huang Y, Linnenkamp U, Guariguata L, Cho N, Cavan D, Shaw J, Makaroff L (2017) IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract 128:40–50. https://doi.org/10.1016/j.diabres.2017.03.024

156. Pan WT (2012) A new fruit fly optimization algorithm: taking the financial distress model as an example. Knowl-Based Syst 26:69–74. https://doi.org/10.1016/j.knosys.2011.07.001

157. Pan Y, Ge SS, Al Mamun A (2009) Weighted locally linear embedding for dimension reduction. Pattern Recogn 42(5):798–811

158. Patel J, TejalUpadhyay D, Patel S (2015) Heart disease prediction using machine learning and data mining technique. Heart Dis 7(1):129–137

159. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, Abu-Hanna A (2009) The coming of age of artificial intelligence in medicine. Artif Intell Med 46(1):5–17

160. Patil B, Joshi R, Toshniwal D (2010) Hybrid prediction model for type-2 diabetic patients. Expert Syst Appl 37(12):8102–8108. https://doi.org/10.1016/j.eswa.2010.05.078

161. Patra R (2020) Prediction of lung cancer using machine learning classifier. In: International Conference on Computing Science, Communication and Security, Springer, pp 132–142. https://doi.org/10.1007/978-981-15-6648-6_11

162. Paul A, Pramanik R, Malakar S, Sarkar R (2021) An ensemble of deep transfer learning models for handwritten music symbol recognition. Neural Comput Appl. https://doi.org/10.1007/s00521-021-06629-

163. Polakowski W, Cournoyer D, Rogers S, DeSimio M, Ruck D, Hoffmeister J, Raines R (1997) Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency. IEEE Trans Med Imaging 16(6):811–819. https://doi.org/10.1109/42.650877

164. Polat H, Mehr HD, Cetin A (2017) Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J Med Syst. https://doi.org/10.1007/s10916-017-0703-x

165. Prakash PS, Rajkumar N (2020) Improved local fisher discriminant analysis based dimensionality reduction for cancer disease prediction. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02542-6

166. Preetha R, Jinny SV (2020) Early diagnose breast cancer with pca-lda based fer and neuro-fuzzy classification system. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02395-z

167. Pujianto U, Ramadhani NA, Wibawa AP (2018) Support vector machine with purified k-means clusters for chronic kidney disease detection. In: 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), IEEE, pp 56–60. https://doi.org/10.1109/eiconcit.2018.8878511

168. Purwar A, Singh SK (2015) Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 42(13):5621–5631. https://doi.org/10.1016/j.eswa.2015.02.050

169. Radhika P, Nair RA, Veena G (2019) A comparative study of lung cancer detection using machine learning algorithms. In: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, pp 1–4. https://doi.org/10.1109/ICECCT.2019.8869001

170. Rahman A, Hasan MM (2018) Automatic detection of white blood cells from microscopic images for malignancy classification of acute lymphoblastic leukemia. In: 2018 International Conference on Innovation in Engineering and Technology (ICIET), IEEE, pp 1–6. https://doi.org/10.1109/ciet.2018.8660914

171. Raja P, Thangavel K (2020) Missing value imputation using unsupervised machine learning techniques. Soft Comput 24(6):4361–4392

172. Rakhlin A, Shvets A, Iglovikov V, Kalinin AA (2018) Deep convolutional neural networks for breast cancer histology image analysis. In: Lecture Notes in Computer Science, Springer International Publishing, pp 737–744. https://doi.org/10.1007/978-3-319-93000-8_83

173. Rakshit S, Manna S, Biswas S, Kundu R, Gupta P, Maitra S, Barman S (2017) Prediction of diabetes type-II using a two-class neural network. In: Communications in Computer and Information Science. Springer, Singapore, pp 65–71. https://doi.org/10.1007/978-981-10-6430-2_6

174. Ramana BV, Boddu RSK (2019) Performance comparison of classification algorithms on medical datasets. In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, pp 0140–0145. https://doi.org/10.1109/ccwc.2019.8666497

175. Rashmi R, Prasad K, Udupa CBK (2022) Breast histopathological image analysis using image processing techniques for

diagnostic purposes: a methodological review. J Med Syst 46(1):1–24

176. Rawat J, Singh A, Bhadauria HS, Virmani J, Devgun JS (2017) Classification of acute lymphoblastic Leukaemia using hybrid hierarchical classifiers. Multimed Tools Appl 76(18):19057–19085. https://doi.org/10.1007/s11042-017-4478-3

177. Rigla M, García-Sáez G, Pons B, Hernando ME (2018) Artificial intelligence methodologies and their application to diabetes. J Diabetes Sci Technol 12(2):303–310

178. Roy K, Banik D, Bhattacharjee D, Nasipuri M (2019) Patch-based system for classification of breast histology images using deep learning. Comput Med Imaging Graphics 71:90–103

179. Roy S, Sarkar D, Malakar S, Sarkar R (2021) Offline signature verification system: a graph neural network based approach. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-021-03592-0

180. Safiyari A, Javidan R (2017) Predicting lung cancer survivability using ensemble learning methods. In: 2017 Intelligent Systems Conference (IntelliSys), IEEE, pp 684–688. https://doi.org/10.1109/IntelliSys.2017.8324368

181. Sahlol AT, Kollmannsberger P, Ewees AA (2020) Efficient classification of white blood cell leukemia with improved swarm optimization of deep features. Sci Rep 10(1):1–11

182. Saji SA, Balachandran K (2015) Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction. In: 2015 International Conference on Advances in Computer Engineering and Applications, IEEE, pp 201–206. https://doi.org/10.1109/icacea.2015.7164695

183. Salaken SM, Khosravi A, Khatami A, Nahavandi S, Hosen MA (2017) Lung cancer classification using deep learned features on low population dataset. In: 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, pp 1–5. https://doi.org/10.1109/CCECE.2017.7946700

184. Samko O, Marshall AD, Rosin PL (2006) Selection of the optimal parameter value for the isomap algorithm. Pattern Recogn Lett 27(9):968–979

185. Sánchez-Cauce R, Pérez-Martín J, Luque M (2021) Multi-input convolutional neural network for breast cancer detection using thermal images and clinical data. Comput Methods Programs Biomed 204:106045

186. Sandhiya S, Palani U (2020) An effective disease prediction system using incremental feature selection and temporal convolutional neural network. J Ambient Intell Hum Comput 11(11):5547–5560

187. Santhakumar D, Logeswari S (2020) Efficient attribute selection technique for leukaemia prediction using microarray gene data. Soft Comput. https://doi.org/10.1007/s00500-020-04793-z

188. Santhakumar D, Logeswari S (2021) Hybrid ant lion mutated ant colony optimizer technique for leukemia prediction using microarray gene data. J Ambient Intell Hum Comput 12(2):2965–2973

189. Sanyal R, Kar D, Sarkar R (2021) Carcinoma type classification from high-resolution breast microscopy images using a hybrid ensemble of deep convolutional features and gradient boosting trees classifiers. IEEE/ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBB.2021.3071022

190. Sedighi Z, Ebrahimpour-Komleh H, Mousavirad SJ (2015) Feature selection effects on kidney desease analysis. In: 2015 International Congress on Technology, Communication and Knowledge (ICTCK), IEEE, pp 455–459. https://doi.org/10.1109/ictck.2015.7582712

191. Sefidian AM, Daneshpour N (2019) Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. Expert Syst Appl 115:68–94

192. Sekh AA, Opstad IS, Birgisdottir AB, Myrmel T, Ahluwalia BS, Agarwal K, Prasad DK (2020) Learning nanoscale motion patterns of vesicles in living cells. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14014–14023

193. Sekh AA, Opstad IS, Godtliebsen G, Birgisdottir ÅB, Ahluwalia BS, Agarwal K, Prasad DK (2021) Physics-based machine learning for subcellular segmentation in living cells. Nat Mach Intell 3(12):1071–1080

194. Sen SK (2017) Predicting and diagnosing of heart disease using machine learning algorithms. Int J Eng Comput Sci. https://doi.org/10.18535/ijecs/v6i6.14

195. Sethy PK, Behera SK (2022) Automatic classification with concatenation of deep and handcrafted features of histological images for breast carcinoma diagnosis. Multimed Tools Appl. https://doi.org/10.1007/s11042-021-11756-5

196. Shafique S, Tehsin S (2018) Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. Technol Cancer Res Treat 17:153303381880278. https://doi.org/10.1177/1533033818802789

197. Shafique S, Tehsin S, Anas S, Masud F (2019) Computer-assisted acute lymphoblastic leukemia detection and diagnosis. In: 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE), IEEE, pp 184–189. https://doi.org/10.1109/c-code.2019.8680972

198. Shakeel PM, Burhanuddin MA, Desa MI (2019) Lung cancer detection from ct image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement 145:702–712

199. Shallu Mehra R (2018) Breast cancer histology images classification: training from scratch or transfer learning? ICT Express 4(4):247–254. https://doi.org/10.1016/j.icte.2018.10.007

200. Shao YE, Hou CD, Chiu CC (2014) Hybrid intelligent modeling schemes for heart disease classification. Appl Soft Comput 14:47–52. https://doi.org/10.1016/j.asoc.2013.09.020

201. Sharma A, Rani R (2021) A systematic review of applications of machine learning in cancer prediction and diagnosis. Arch Comput Methods Eng 28(7):4875–4896

202. Shaw SS, Ahmed S, Malakar S, Garcia-Hernandez L, Abraham A, Sarkar R (2021) Hybridization of ring theory-based evolutionary algorithm and particle swarm optimization to solve class imbalance problem. Complex Intell Syst 7(4):2069–2091

203. Shaw SS, Ahmed S, Malakar S, Sarkar R (2021b) An ensemble approach for handling class imbalanced disease datasets. In: Proceedings of International Conference on Machine Intelligence and Data Science Applications, Springer, pp 345–355

204. Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, Yang B, Liu D (2016) Evolving support vector machines using fruit fly optimization for medical data classification. Knowl-Based Syst 96:61–75. https://doi.org/10.1016/j.knosys.2016.01.002

205. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

206. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

207. Sisodia DS, Verma A (2017) Prediction performance of individual and ensemble learners for chronic kidney disease. In: 2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE, pp 1027–1031. https://doi.org/10.1109/icici.2017.8365295

208. Sonawane JS, Patil D (2014) Prediction of heart disease using multilayer perceptron neural network. In: International Conference on Information Communication and Embedded Systems (ICICES2014), IEEE, pp 1–6. https://doi.org/10.1109/ICICES.2014.7033860

209. Sridhar K, Baskar S, Shakeel PM, Dhulipala VS (2019) Developing brain abnormality recognize system using multi-objective pattern producing neural network. J Ambient Intell Humaniz Comput 10(8):3287–3295

210. Srivenkatesh M (2019) Performance evolution of different machine learning algorithms for prediction of liver disease. International Journal of Innovative Technology and Exploring Engineering 9(2), 1115–1122, https://doi.org/10.35940/ijitee.l3619.129219

211. Stöger K, Schneeberger D, Holzinger A (2021) Medical artificial intelligence: the European legal perspective. Commun ACM 64(11):34–36

212. Sudharshan P, Petitjean C, Spanhol F, Oliveira LE, Heutte L, Honeine P (2019) Multiple instance learning for histopathological breast cancer image classification. Expert Syst Appl 117:103–111. https://doi.org/10.1016/j.eswa.2018.09.049

213. Sumi TA, Hossain MS, Andersson K (2022) Automated Acute Lymphocytic Leukemia (ALL) Detection Using Microscopic Images: An Efficient CAD Approach. In: Proceedings of Trends in Electronics and Health Informatics, Springer, pp 363–376

214. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–9

215. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2818–2826

216. Taher F, Werghi N, Al-Ahmad H (2015) Rule based classification of sputum images for early lung cancer detection. In: 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS), IEEE, pp 29–32. https://doi.org/10.1109/icecs.2015.7440241

217. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X (2020) Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med Image Anal 63:101693

218. Tekade R, Rajeswari K (2018) Lung cancer detection and classification using deep learning. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, pp 1–5. https://doi.org/10.1109/iccubea.2018.8697352

219. Ting KM, Zheng Z (2003) A study of adaboost with Naive Bayesian classifiers: weakness and improvement. Comput Intell 19(2):186–200

220. Tiwari M, Chakrabarti P, Chakrabarti T (2018) Performance analysis and error evaluation towards the liver cancer diagnosis using lazy classifiers for ILPD. In: Soft Computing Systems, Springer Singapore, pp 161–168. https://doi.org/10.1007/978-981-13-1936-5_19

221. Vaishali R, Sasikala R, Ramasubbareddy S, Remya S, Nalluri S (2017) Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pima indians diabetes dataset. In: 2017 International Conference on Computing Networking and Informatics (ICCNI), IEEE, pp 1–5. https://doi.org/10.1109/iccni.2017.8123815

222. Vas M, Dessai A (2017) Lung cancer detection system using lung CT image processing. In: 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE, pp 1–5. https://doi.org/10.1109/iccubea.2017.8463851

223. Vashisth S, Dhall I, Saraswat S (2020) Chronic kidney disease (CKD) diagnosis using multi-layer perceptron classifier. In: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, pp 346–350. https://doi.org/10.1109/confluence47617.2020.9058178

224. Wei L, Su R, Wang B, Li X, Zou Q, Gao X (2019) Integration of deep feature representations and handcrafted features to improve the prediction of n6-methyladenosine sites. Neurocomputing 324:3–9

225. Wibawa HA, Malik I, Bahtiar N (2018) Evaluation of kernel-based extreme learning machine performance for prediction of chronic kidney disease. In: 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), IEEE, pp 1–4. https://doi.org/10.1109/icicos.2018.8621762

226. Wibawa MS, Maysanjaya IMD, Putra IMAW (2017) Boosted classifier and features selection for enhancing chronic kidney disease diagnose. In: 2017 5th International Conference on Cyber and IT Service Management (CITSM), IEEE, pp 1–6. https://doi.org/10.1109/citsm.2017.8089245

227. Wu H, Yang S, Huang Z, He J, Wang X (2018) Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 10:100–107. https://doi.org/10.1016/j.imu.2017.12.006

228. Yan R, Ren F, Wang Z, Wang L, Zhang T, Liu Y, Rao X, Zheng C, Zhang F (2020) Breast cancer histopathological image classification using a hybrid deep neural network. Methods 173:52–60. https://doi.org/10.1016/j.ymeth.2019.06.014

229. Yang D, Xu Z, Li W, Myronenko A, Roth HR, Harmon S, Xu S, Turkbey B, Turkbey E, Wang X et al (2021) Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy. Japan. Med Image Anal 70:101992

230. Yang Z, Ran L, Zhang S, Xia Y, Zhang Y (2019) EMS-net: ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. Neurocomputing 366:46–53. https://doi.org/10.1016/j.neucom.2019.07.080

231. Zhang Q, Kong X (2020) Design of automatic lung nodule detection system based on multi-scene deep learning framework. IEEE Access 8:90380–90389

232. Sq Zhang (2009) Enhanced supervised locally linear embedding. Pattern Recogn Lett 30(13):1208–1218

233. Zhang Y, Zhou B, Cai X, Guo W, Ding X, Yuan X (2021) Missing value imputation in multivariate time series with end-to-end generative adversarial networks. Inf Sci 551:67–82

234. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Syst Appl 41(4):1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

235. Zhou T, Ruan S, Canu S (2019) A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3:100004

236. Zhou Y, Lu Y, Pei Z (2021) Accurate diagnosis of early lung cancer based on the convolutional neural network model of the embedded medical system. Microprocess Microsyst 81:103754

237. Zhu C, Idemudia CU, Feng W (2019) Improved logistic regression model for diabetes prediction by integrating PCA and k-means techniques. Inform Med Unlocked 17:100179. https://doi.org/10.1016/j.imu.2019.100179