Journal of Electrical Systems and Information Technology

REVIEW **Open Access**

A comprehensive review for chronic disease prediction using machine learning algorithms

Rakibul Islam[1*], Azrin Sultana[1] and Mohammad Rashedul Islam[2,3]

*Correspondence:
rakibulislam.cse21@gmail.com

[1] Department of Computer Science, American International University-Bangladesh, Dhaka 1229, Bangladesh
[2] Department of Research and Training Monitoring, Bangladesh College of Physicians and Surgeons, Dhaka 1212, Bangladesh
[3] Department of Health Informatics, Bangladesh University of Health Sciences, Dhaka 1216, Bangladesh

Abstract

The past few years have seen an emergence of interest in examining the significance of machine learning (ML) in the medical field. Diseases, health emergencies, and medical disorders may now be identified with greater accuracy because of technological advancements and advances in ML. It is essential especially to diagnose individuals with chronic diseases (CD) as early as possible. Our study has focused on analyzing ML's applicability to predict CD, including cardiovascular disease, diabetes, cancer, liver, and neurological disorders. This study offered a high-level summary of the previous research on ML-based approaches for predicting CD and some instances of their applications. To wrap things up, we compared the results obtained by various studies and the methodologies as well as tools employed by the researchers. The factors or parameters that are responsible for improving the accuracy of the predicting model for different previous works are also identified. For identifying significant features, most of the authors employed a variety of strategies, where least absolute shrinkage and selection (LASSO), minimal-redundancy-maximum-relevance (mRMR), and RELIEF are extensively used methods. It is seen that a wide range of ML approaches, including support vector machine (SVM), random forest (RF), decision tree (DT), naïve Bayes (NB), etc., have been widely used. Also, several deep learning techniques and hybrid models are employed to create CD prediction models, resulting in efficient and reliable clinical decision-making models. For the benefit of the whole healthcare system, we have also offered our suggestions for enhancing the prediction results of CD.

Keywords: Machine learning, Chronic disease prediction, Artificial intelligence, Machine learning in healthcare, Data mining, Heart disease prediction, Diabetes disease

Introduction

In the last 20 years, machine learning (ML) has advanced considerably from being a research curiosity to a useful technology with widespread commercial applications. It is a branch of artificial intelligence (AI) that employs statistical methods to fit models to data and discover relevant patterns from massive, unstructured, and complicated datasets [1]. It is a comprehensive, multidisciplinary field with roots in statistics, mathematics, computer science, and cognitive analytics, among other disciplines [2]. Algorithms trained by ML systems can utilize past data to make accurate predictions about unseen data. The basis of the ML process is observations of data, such as examples, firsthand knowledge, or instructions. It searches for patterns in the data to subsequently draw conclusions from the supplied instances. The main goal of ML is to make it possible for computers to learn independently, without human aid, and to adapt after retraining. To predict future outputs, the supervised ML algorithm trains a model using historical data on both inputs and outputs, whereas unsupervised ML explores intrinsic structures and hidden patterns in input data [3–5].

ML approaches have recently had a considerable impact on the healthcare industry (HI). The use of ML techniques in healthcare can lead to advancements such as more precise prediction models, new treatment approaches, clinical decision support systems (CDSS), medication development, and reductions in healthcare expenditures [6, 7]. Recent practical uses of ML in healthcare have been enabled by the collection of daily healthcare data as well as the advancement of big data processing. Different ML techniques can be applied to those datasets, which may be in structured or unstructured form, to provide a better outcome in healthcare. Various ML algorithms, such as linear regression (LR), support vector machine (SVM), random forest (RF), decision tree (DT), K-nearest neighbor (KNN), deep learning (DL), artificial neural network (ANN), and boosting algorithms are widely used to predict diseases [8, 9]. Using ML algorithms to forecast which treatment protocols would work best for a particular patient based on their characteristics and the state of the treatment is known as a method of ML in the HI. ML applications require a training dataset that includes an outcome variable for building various models for physicians and patients [10].

Chronic disease (CD) is a condition or illness that lasts for at least three months and can have serious long-term consequences. CD is more common in elderly people and can typically be managed but not cured [11, 12]. Cancer, cardiovascular disease (CVD), diabetes, brain disease, liver disease, stroke, and arthritis are common forms of CD [13]. The World Health Organization (WHO) estimates that CD causes 41 million deaths annually, or 74% of all deaths worldwide. Each year, 17 million people under the age of 70 die from a CD; however, only 15% of these premature deaths occur in countries with high incomes [14, 15]. CVD causes the most significant number of CD deaths, followed by cancer and diabetes. Smoking, lack of exercise, excessive alcohol consumption, and poor nutrition contribute to an increased risk of dying from a CD [16]. In the field of healthcare informatics, CD prediction plays a significant role. CD diagnosis systems can be very effective in correctly scheming and taking care of CD patients [17, 18]. The only way to reduce mortality and prepare for future diseases is to predict them early so that patients can receive proper treatment and disease severity can be prevented [19]. Patients require a disease prediction model with the help of various supervised ML algorithms such as RF, DT, KNN, ANN, NN, SVM, NLP, and many more, allowing health officials and doctors to take preventative measures that can reliably, accurately, and efficiently predict diseases [20, 21].

Although numerous works have been conducted on individual CDs, where most of the researchers have discussed different aspects and outcomes of that specific disease, we have tried to bring all the CDs under the same umbrella. Therefore, the aim of this systemic review is to provide a comprehensive overview of the previous studies regarding the predicting model of different CDs, in which we give more emphasis on representing comparative tabular data based on previous research so that readers can easily know about the description of the dataset, findings, outcomes and different key factors which helped to improve the accuracy of their proposed system. Furthermore, we have also provided the list of datasets available for the classification of different chronic diseases.

Contribution of this study

The main contributions of this study are as follows:

- This study focuses on how ML algorithms are used to predict CDs such as liver disease, cancer disease, brain disease, heart disease, and diabetes.
- This article covered the author's proposed system and findings, objectives, data sources, technologies, algorithms, and the accuracy of their study.
- This study also addressed the future direction for cost-effective medical care by integrating the predictive model (PM) into the healthcare system.
- This comprehensive review of different CD predictions can be helpful for future researchers.

The remaining sections of this study are briefly arranged as follows: Sect. "Methodology" discusses the entire journey of paper selection and review from various journals. Sect. "Predictive model (PM) using ML algorithms" provides short information on how PM works; in Sect. "ML for CD prediction", previous studies have been reviewed where ML is used for predicting and diagnosing different CDs. The remaining sections cover the discussion and conclusion, respectively.

Methodology

We mainly looked for no articles using high-impact factors publisher databases such as Wiley Oxford journals, The Lancet, Springer, IEEE, Hindawi, ACM, and ScienceDirect. As shown in Fig. 1, more than 470 papers were screened for our investigation, and the search titles for the papers were “ML in healthcare,” “chronic disease prediction,” and “chronic disease classification.” The entire paper collection or searching process consisted of two steps. As this study worked with prediction models of CD, in the first phase, chronic disease prediction or classification papers were searched. And in the second phase, the applicability of the paper to the study was thoroughly scrutinized. Most of these research articles selected for CD prediction were released between 2018 and 2024. In addition, this study only examined papers with a high number of citations (more than ten) or a relevant abstract and title for further investigation. During this process, articles were included in the collection only if all the writers deemed them appropriate; any differences were resolved by consensus. Thus, 125 papers out of 473 were discovered that were pertinent to our investigation.

Predictive model (PM) using ML algorithms

PM can predict future outcomes by assessing past results and existing data. PM has gradually integrated into data mining using AI technologies and ML algorithms. It has improved the quality of decision-making processes and allowed for better foresight into potential outcomes [22, 23]. PM consists of seven phases (Fig. 2), beginning with the

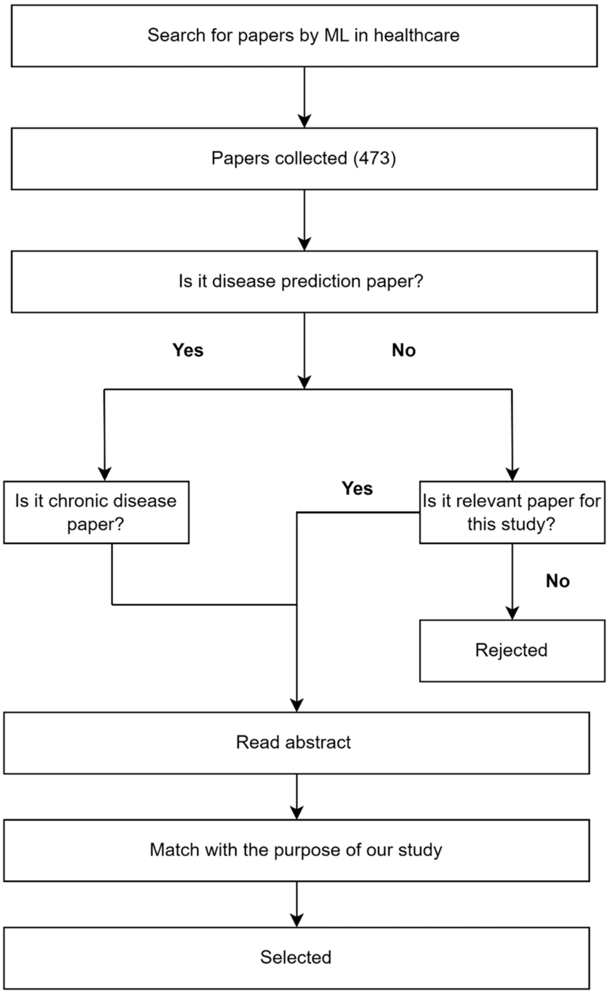
Search for papers by ML in healthcare
Papers collected (473)
Is it disease prediction paper?
Yes
No
Is it chronic disease paper?
Yes
Is it relevant paper for this study?
No
Rejected
Read abstract
Match with the purpose of our study
Selected

Fig. 1 Flow diagram of the paper selection process

data collection process. Data collection is followed by data preparation. Data preparation included removing missing values, scaling, eliminating outliers, and balancing the dataset. Selecting the ML model is the third step in this process. After the ML model has been chosen, the dataset should be split. The fifth stage, which involves training the dataset with the selected model, is the most important. Many researchers use hyperparameter tuning techniques to improve accuracy [24, 25]. Not only on HI but predictive analytic techniques and tools may also now reliably foretell a company's sales and profit future. This is because the PM now incorporates sales [26]. Another area is marketing, explicitly anticipating customers' reactions and needs based on information gleaned from feedback [27]. Social media is the industry that enables platforms to identify client behavior and predict future consequences [28]. As is seen, PM has several uses, but one of the most important is risk assessment, which may assess risk and ascertain the degree of profit or loss that the future contains. Predictive analytics for quality improvement consider previous comments, adjustments, and suggestions that might improve the quality [29, 30]. The prediction model is widely utilized in the HI since it has become a valuable tool for making medical decisions as patients react differently to every form of

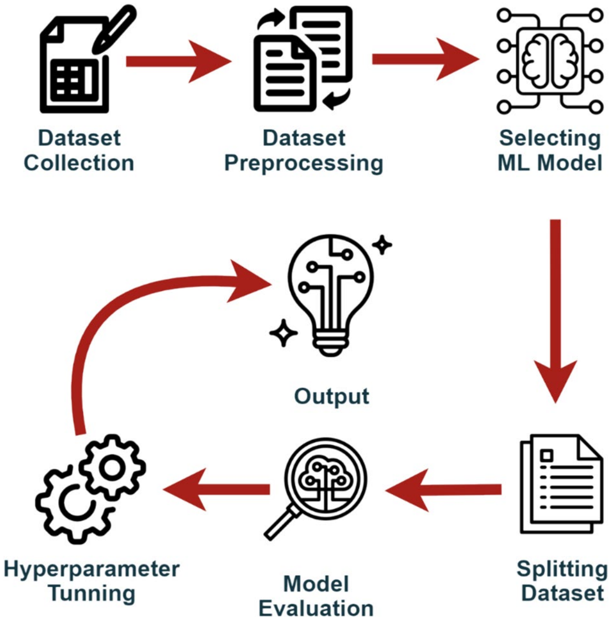

Fig. 2 Prediction model flow diagram

treatment, particularly for chronic diseases [19]. An opportunity to develop a practical preventative and treatment approach can be presented by an early diagnosis. ML models are used to predict disease, which helps doctors categorize high-risk patients, provide a unique diagnosis, reduce risk, and eliminate health hazards [31, 32]. This research mainly concentrated on reviewing predicting models of CD.

ML for CD prediction

Liver disease (LD)

After the epidermis, the liver is the biggest internal organ. In terms of size and location, it is roughly the size of a football and rests just beneath the right ribs. As food travels through the digestive system, the liver sorts out the good from the bad. It also secretes bile, which helps in digestion and eliminates harmful substances from the body [33]. The liver's ability to operate correctly declines when scar tissue gradually replaces healthy liver tissue. LD can potentially lead to patient to liver failure and cancer if not addressed on time. There are approximately 2 million deaths a year caused by LD, 1 million from cirrhosis, and another 1 million from hepatocellular carcinoma and viral hepatitis [34, 35]. LD has several potential causes, including infections, poor dietary choices, drug use, alcohol abuse, and toxic exposure. Genetic predispositions to LD exist as well. Hepatitis A, B, C, D, & E are all forms of viral hepatitis, while fatty LD results from poor dietary choices and an unhealthy lifestyle. Hepatitis B and C are the most prevalent of these five forms of the disease. Every 30 s, a new hepatitis patient dies, and 11% of the world's population succumbs to the disease each year. It is found that between 20 and 40% of the population in Western industrialized countries has nonalcoholic fatty LD (NAFLD). Along with rising rates of obesity, T2D, and metabolic syndrome, NAFLD has been on the rise in recent years [36, 37]. The construction of more precise prediction models employing a wide range of ML methods is becoming increasingly popular in response to the expanding use of ML in the healthcare industry. Early prediction of associated risk factors of LD could help a lot with diagnosis, prevention, or treatment [38]. Therefore, this study aims to analyze previous research on the LD prediction model by providing information about the dataset, research objective, algorithms used, findings and different important aspects of their study (Table 1).

Liu et al. [39] used seven ML algorithms to predict Non-alcoholic fatty liver disease using a dataset of 15,315 cases and 35 characteristics from the International Health Care Center. BMI was the most significant indication based on the feature ranking. The dataset was partitioned at random in a 7:3 ratio. This study used the Tensor flow framework to create the multilayer perceptron (MLP), CNN, and long short-term memory networks (LSTM) models. At the same time, the Python scikit-learn library was accountable for developing the XGBoost, SVM, Stochastic gradient descent (SGD) classifier, and LR model. Model efficacy was evaluated using nine different matrices. In comparison to all metrics, XGBoost offers the best accuracy.

Liu et al. [40] built a prediction model for liver patients to predict the recurrence risk of hepatocellular carcinoma patients. Additionally, they constructed a web-based personal assessment system for the patient. The proposed approaches utilized six ML algorithms. The dataset sample size was 315. The study was conducted by splitting the dataset as a ratio of 7:3 for training and testing, respectively. The author applied Synthetic Minority Over-sampling Technique (SMOTE) and replaced missing values in the pre-processing stage. MLP obtained the highest accuracy in this research.

Cao et al. [41] suggested a technique to predict and evaluate NAFLD patients with several ML methods. Four distinct models were developed for making predictions, and their performance was compared to determine the most suitable model. The sample size of their study was 22,140. Out of the four ML algorithms analyzed, the XGBoost model exhibited the best results with the subsequent metrics: accuracy (83.5%), specificity (83.4%), sensitivity (83.5%), Youden index (66.9%), recall (83.5%), precision (83.1%), F-1 score (83.3%), and AUC (91.4%).

According to Harrison et al. [42], the near-term mortality of patients with liver cirrhosis was predicted using the two ML algorithms LR and LTSM. Their study aimed to integrate the suggested model into an electronic health record system with the aim of facilitating precise and prompt forecasts of decompensation and mortality. This PM used the dataset consisting of 62 features and 340,553 records from the ICU at Virginia Health System. The effectiveness and generalizability of each model were verified by testing them on an anonymous data set comprising information on the 2017 patient stay.

Speiser et al. [43] predicted the daily status of patients with acute liver failure brought on by acetaminophen use. They assessed the effectiveness of methods for outcomes in the first week of hospitalization. The acute Liver Failure Study Group (ALFSG) database served as the source of information, and its sample size of 1042 included 14 characteristics. Generalized linear mixed models, Bayesian GLMM, binary mixed model tree and forest, were among the methods. RF, SVM, KNN,

Table 1 Overview of different parameters from the previous works on LD

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	Validation	ML algorithms	Findings
Liu et al. [39]	2021	Detect Non-alcoholic fatty liver disease	International Health Care Center	15,315	35	BMI, WC, triglyceride, WHtR, ALT, weight, and FBG	N/A	XGBoost, LR, SVM, SGD, MLP, CNN, and LSTM	84.6% accuracy using XGBoost
Liu et al. [40]	2023	Predict recurrence risk for hepatocellular carcinoma patients	Third Affiliated Hospital of Sun Yat-sen University	315	13	GGT, fibrinogen, neutrophil, AST and TB	tenfold	AB, LR, MLP, BAG, GBM, XGBoost	The highest accuracy of 62.7% using MLP
Cao et al. [41]	2024	Evaluate potential NAFLD patients	Beijing Health Management Cohort	22,140	16	AST, CMI, BMI, ALT, and TyG index	tenfold	DT, NB, LR, and XGBoost	Highest 83.5% accuracy using XGBoost
Harrison et al.[42]	2018	Predict the Near-Term Mortality of the liver Cirrhosis Patients	University of Virginia Health System	340,553	62	SI, WBCC, SPO2, BV, PCO2, RBCC and 16 more	sevenfold	LR, RNN-LSTM	Highest accuracy 84% using LR
Speiser et al. [43]	2019	Predicting the daily outcome of the patients with acetaminophen-induced acute liver failure	ALFSG database	4461	16	lactate, ammonia, ALT, CREAT, bilirubin, AST, phosphate, platelets, and age	N/A	RF, SVM, ANN, Frequentist GLMM, Bayesian GLMM, BiMM Tree, CART, and BiMM forest	Highest accuracy 92% using e BiMM tree
Safdari et al. [44]	2022	Predict hepatitis C	From UCI repository	615	13	N/A	fivefold	DT, KNN, RF, SVM, NB, and LR	Highest accuracy of 97.29% using RF
Goldar et al. [45]	2020	Diagnosis of Non-Alcoholic Fatty Liver	Iranian liver specialists	400	7	SGOT, SGPT, ALP, FBS, Platelet, weight, and Cholesterol	N/A	ANFIS-PSO	RMSE value 0.3712
Barus et al.[46]	2022	Predict Liver Disease	ILPD	583	11	N/A	N/A	LR and SVM	The highest accuracy of 87% using SVM
Feldman et al. [47]	2021	Prediction of antiviral therapy duration for patients with hepatitis C	Privately collected	6775	14	CC, NRD, DC, age, HC	fivefold	XGBoost, SVM, and RF	Older patients with Compensated or decompensated cirrhosis, and also with hepatocellular carcinoma, receive prolonged DAA therapy
Gupta et al. [48]	2022	Liver disease prediction	ILPD	583	11	N/A	N/A	LR, NB, DT, RF, GB, LightGB, AdaBoost, Extreme GB, KNN, and Stacking	Highest accuracy of 63% using RF

Table 1 (continued)

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	Validation	ML algorithms	Findings
Ahad et al. [49]	2024	Hepatitis C prediction and built a user interface	UCI repository	615	12	Age, sex, ALP, ALT, Albumin, Bilirubin, CREAT etc	fivefold	RF, SVM, XGBoost, DT, LR	Highest accuracy of 99.84 with Ensemble LR

Alkphos, alkaline phosphatase; ALT, alanine aminotransferase; AST, Aspartate Aminotransferase; BAG, bootstrapped aggregating; BMI, body mass index; BV, Bicarbonate Value; CC, Compensated cirrhosis; CMI, cardiometabolic index; CREAT, creatinine; DAA, Direct Acting Antiviral; DB, direct bilirubin; DC, decompensated cirrhosis; FBG, fasting blood glucose; FBS, fasting blood sugar; HC, Hepatocellular carcinoma; N/A, missing or not available; NB, naïve Bayes; NRD, Number of relevant diseases; PC, Platelet count; PCO2, Partial Pressure of Carbon Dioxide in the Blood; RBCC, Red Blood Cell Count; RMSE, Root Mean Square Error; SGOT, serum glutamic-oxaloacetic transaminase; SGPT, serum glutamate pyruvate transaminase; SI, Shock Index, WBCC, White Blood Cell Count; SPO2, Peripheral Capillary Oxygen Saturation; TB, total bilirubin; triglyceride-glucose index, TyG index; WC, waist circumference; WHtR, waist-to-height ratio; γ-glutamyl transpeptidase, GGT

ANN, and CART were also applied. Utilizing ROC, sensitivity, and specificity, the model was verified. BiMM trees gave the maximum accuracy level for this PM.

Cancer

Cancer is a condition characterized by the proliferation of abnormal cells that have the ability to invade or spread to other regions of the body [50]. There are more than two hundred distinct varieties of cancer, and they can be categorized based on their location or starting point within the body. Most cancer-related diseases and fatalities result from tumor cells reemerging in nearby organs and tissues [51]. Malnutrition in cancer is probably caused by more than one thing, but the location of the tumor and the symptoms that show up, such as anorexia, taste changes, dysphagia, nausea, vomiting, and diarrhea, can make nutrition and functional ability even worse [52, 53]. As per the findings of Global Cancer Statistics, lung cancer constituted the leading cause of cancer-related mortality (1.8 million deaths, 18%), followed by colorectal cancer (CRC) (9.4%) and liver cancer (8.3%). Among all cancer types, breast cancer accounted for the highest number of newly identified cases (2.3 million), representing 11.7% of all, followed by lung cancer, which is 11.4% and CRC (10%) [54]. This study includes skin cancer, ovarian cancer, breast cancer, gastric cancer, lung cancer and thyroid cancer. Cancer research has shifted its focus to early detection and prognosis because of the positive impact it may have on the clinical care of patients [55, 56]. Several studies have been conducted to create an efficient predictive model for cancer patients. Therefore, this study aims to analyze previous research on cancer disease prediction by providing information about the dataset, research objective, algorithms used, findings and different important aspects of their study (Table 2).

Abbasi et al. [57] predicted skin cancer by employing the Kaplan–Meier estimator and Cox proportional hazards regression model, utilizing eight ML classifiers on a publicly available dataset from the ICGC Data Portal, specifically targeting skin cutaneous melanoma cancers. Additionally, four different ensemble methods (stacking, bagging, boosting, and voting) were created and trained to achieve optimal results. The performance was evaluated and interpreted using accuracy, precision, recall, F1 score, confusion matrix, and ROC curves, the RF classifier achieved an outstanding accuracy of 99%.

Using an image pre-processing technique to filter and eliminate the excess noise existing in the picture by various approaches, Murugan et al. [59] suggested a methodology to predict skin cancer. The median filter is used to determine the location of the skin region of the affected area, and the mean shift segmentation technique was then utilized to divide the afflicted area from the surrounding healthy skin. SVM, probabilistic neural networks (PNN), RF, and Combined SVM + RF classifiers have all been employed as the methods for this study. Compared to other classifiers, the results produced by the combined SVM + RF classifier were better. The total number of images used in the experiment is 1000, and 10 cross-validation was used, with all samples trained and tested.

Naji et al. [60] explored the use of ML algorithms to predict breast cancer and determine which algorithms were most efficient in terms of accuracy, precision, and confusion matrix. This article primarily compares the effectiveness of five classifiers: SVM, RF,

Table 2 Overview of different parameters from the previous works on cancer disease

Author	Year	Cancer type	Objective	Dataset source	Dataset instance	Dataset features	Validation	ML methods	Findings
Abbasi et al. [57]	2023	Skin cancer	Predict skin cancer survival rate	ICGC Data Portal	1,797,138	31	N/A	RF, DT, GB, AB, GNB, ET, LR, and Light GBM	The highest accuracy is 99% using RF
Sheela Lavanya et al. [58]	2024	Ovarian cancer	Predict early stage of ovarian cancer	Kaggle	349	51	Fivefold	SVM, DT, KNN, GB, XGB, MVC, and Stacking	The highest accuracy is 89%
Murugan et al. [59]	2021	Skin cancer	Predict skin cancer	Kaggle	1000	N/A	Tenfold	SVM, PNN, RF, and Combined SVM + RF	Highest Accuracy 89.31% using SVM + RF
Naji et al. [60]	2021	Breast cancer	Predict breast cancer	Wisconsin dataset	569	N/A	N/A	SVM, RF, LR, C4.5 and KNN	Highest accuracy of 97.2% using SVM
Sakai et al. [61]	2018	Gastric cancer	Predict early stage of gastric cancer using endoscopic images	Used 2 dataset	1000	N/A	N/A	CNN	Accuracy 87.6%
Salmi et al. [62]	2019	Colon cancer	Predict colon cancer	Al-Islam Hospital Bandung	209	7	N/A	NB	Accuracy 95.24%
Hasan et al. [63]	2022	Colon cancer	Predict colon cancer	Privately collected	500	N/A	N/A	DCNN	99.8% accuracy
Adeoye et al. [64]	2022	Oral cancer	Predict oral cavity cancer	Hong Kong Hospital Authority Clinical Management System (HA-CMS)	313	13	fivefold	time-dependent Cox model, DeepSurv, DeepHit, and RSF	Concordance indices 0.94
Xie et al. [65]	2020	Lung cancer	Predict early stage of lung cancer	Hubei Taihe Hospital	110	N/A	tenfold	KNN, NB, AdaBoost, SVM, RF, and Neural Network	The highest accuracy using NB and NN is 100%
Gupta et al. [66]	2019	Colon cancer	Predict colon cancer stages with survival period	Chang Gung Memorial Hospital	4021	21	fivefold	RF, SVM, LR, MLP, KNN, and AB	Accuracy using RF in 84%
Mourad et al. [67]	2020	Thyroid cancer	Predict thyroid cancer	SEER dataset	8477	34	N/A	MLP	Accuracy 94.5%

DCNN, deep convolutional neural network; DeepSurv, deep feed-forward neural network; GNB, Gaussian–naïve Bayes, N/A, missing or not available; RSF, random survival forest; BPNN, back propagation neural network

LR, C4.5, and KNN. 25% of the dataset was utilized for testing, while 75% was used for training. SVM consistently outperformed the other classifiers.

In predicting early stomach cancer in endoscopic pictures, Sakai et al. [61] suggested a convolutional neural network-based automated detection system. The most significant contribution of this study is the effective automated diagnosis of early stomach cancer with weak morphological traits, which might be difficult to identify even for endoscopists. About a thousand white-light imaging pictures of early gastric cancer (particularly kinds 0-I, 0-IIa, and 0-IIc) have been employed in the study. The author retrieved 24-bit full-color pictures with a resolution of 1000 × 870 pixels from the video sequence. The authors collected 172,555 cancer pictures and 176,388 normal images, both measuring 224 × 224 pixels in size, by applying nine different kinds of enhancements, including rotation, shear, shift, flip, and magnification twice. For learning rates of 0.0001 and 0.00001 both before and after 34 epochs, correspondingly, the original network was trained for 50 epochs.

Salmi & Rustam [62] used the NB algorithm to forecast colon cancer, a prediction approach based on a simple probabilistic algorithm with a strong independence assumption. This study's dataset was obtained from Al-Islam Hospital Bandung and consisted of seven columns and 209 instances. Age, Carcinoembryonic Antigen, hemoglobin, leukocytes, hematocrit, and thrombocytes were the features of the dataset. The dataset was divided into 80% and 20% for training and testing. The authors found 95.24% accuracy with the NB algorithm.

Brain disease

The brain is the most significant and complicated human organ responsible for regulating almost every aspect of corporal function. Many neurological ailments, such as Alzheimer's disease (AD), Parkinson's disease (PD), stroke, Meningitis, tumors, cerebral edema, and many more, are eventually due to aging and neuronal death [68–70]. According to a recent study, AD, PD, stroke, epilepsy, migraine, brain traumas, and neuro infections are just some of the many neurological illnesses that affect over a sixth of the global population and claim the lives of about 6.8 million people every year [71, 72]. The neurodegenerative sickness that affects older people most frequently is AD [73]. PD affects 2–4% of the 65 and older population, making it the second most prevalent neurodegenerative condition. Between 4.1 and 4.6 million persons were affected in 2005; experts project that figure would more than double by 2030, reaching 8.7–9.3 million [74]. Regarding diseases that affect the brain, stroke is the most incapacitating long-term ailment [75]. Though men have a higher risk of having an acute stroke at some point in their lives, women have a higher mortality rate from such an event. Therefore, around 16% of all women are expected to die from a stroke, compared with 8% of all males; the discrepancy is primarily owing to the older average age at which strokes occur in women and to the longer average life expectancy of women [76, 77]. Although there have been advancements in surgical and other therapeutic methods, brain disease or brain stroke continues to be one of the leading causes of death and disability. Improving patient quality of life requires accurate and early detection of those with brain diseases. This study aims to cover previous studies of brain disease prediction to analyze the findings, methods and important aspects of their study (Table 3).

Table 3 Overview of different parameters from the previous works on brain disease

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	validation	ML algorithms	Findings
Ranga-Swamy et al. [78]	2020	Predict various variant's effects on AD patients	GWAS and GTE	57,853	39	H3K36me3	Tenfold	RF, XGBoost, AB, and NN	Highest accuracy of 81.21% using RF
Rani et al. [79]	2024	Predict Alzheimer disease using MRI images	Kaggle	416	15	N/A	N/A	RF, DT, and XGBoost	Highest accuracy of 95.03% using RF
Lin et al. [80]	2020	Predict the stroke disease	Taiwan Stroke Registry	58,493	206	discharge NIHSS assessment items, discharge Barthel index, and the 30-day mRS degree	Tenfold	SVM, RF ANN, and HANN	The highest AUC value is 97.1% using SVM
Haq et al. [81]	2019	Diagnosing the PD using voice recording	University of Oxford	195	23	MDVP: Fhi, MDVP: Fo, MDVP: Flo, DFA, HNR, D2, RPDE, spread2 MDVP: Shimmer,, and PPE	Tenfold	SVM	99% accuracy
Kostev et al. [82]	2021	Predict the risk of stroke in patients who are suffering from late-onset epilepsy	IQVIA Disease Analyzer database	11,466	42	diabetes, hypertension, heart failure, and alcohol	N/A	SOMS	Four co-diagnoses in patients with epilepsy were found to be key predictors of stroke in people
Dritsas et al. [83]	2022	Predict stroke risk	Kaggle	3254	11	Age, BMI, Avg_glucose_level	Tenfold	NB, LR, RF, KNN, SGD, DT, Stacking, and MLP	Highest accuracy98% using Scaling
Shoily et al. [84]	2019	Diagnose the stroke disease	Privately collected	1058	28	N/A	Tenfold	NB, J48, KNN, and RF	Highest accuracy 99.8% using j58, KNN, and RF
Murcia et al. [85]	2020	Predicting manifold structure of Alzheimer	ADNI database	479	10	N/A	Tenfold	CAE-SVM and CAE-MLP	Highest accuracy 84.9% using CAE-SVM
Grover et al. [86]	2018	Predicting the severity of PD	Parkinson Tele-monitoring Voice Dataset from UCI repository	5,875	16	N/A	N/A	DNN, and ANFIS-SVR	The highest accuracy of 83.34%, using DNN

Table 3 (continued)

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	validation	ML algorithms	Findings
Sudharsan et al. [87]	2021	Classifying AD of elderly people	Alzheimer's Neuroimaging Initiative database	214	N/A	N/A	Tenfold	RLEM, SVM, and IVM	Highest accuracyc79.03% using SVM
Franciotti et al. [88]	2023	Predict Alzheimer disease from mild cognitive impairment	ADNI database	2000	15	Cognition immediate and delayed verbal memory, and daily functioning	N/A	RF, GB, and XGBoosting	The highest accuracy is 90%

ADNI, Alzheimer's Disease Neuroimaging Initiative; ANFIS, adaptive network-based fuzzy inference system; CAE, Convolutional Autoencoders; DFA, Signal fractal scaling exponent; GTE, Genotype Tissue Expression; GWAS, Genome-Wide Association Study; IVM, Import Vector Machine; MDVP Fhi, Maximum vocal fundamental frequency; MDVP Fo, The average vocal voice fundamental frequency; MDVP Shimmer, Several measures of variation in amplitude; MDVP, Flo, Minimum vocal fundamental frequency; N/A, missing or not available; NCVS, National Center for Voice and Speech; RELM, Regularized Extreme Learning Machine; RPDE, Two nonlinear dynamical complexity measures

A model for anticipating the impact of AD on various variants was suggested in [78]. The study's primary objective was to use ML to create a classification model for estimating the risk that a given variant poses to AD. There are 57,853 instances in all and 39 attributes to analyze. The recursive feature elimination via cross-validation score was utilized to choose the most relevant features. To anticipate the system's accuracy, the authors utilized a variety of ML methods, including RF, XGBoost, AB, and NN, to train and then test the model. Additionally, in this research, the authors developed a web server to find potentially harmful variations linked to AD. Input versions were assigned a score between 0 and 1 by the algorithm. The threshold for determining deleteriousness was 0.38; below that number, a variant is considered harmless.

The proposed framework by Lin et al. [80] was about to predict the outcome of a 90-day stroke using several ML algorithms. The 58,493 data and 206 characteristics from the Taiwan Stroke Registry were used for this analysis. To ensure accurate results, the authors implemented an evaluation validation into the data preparation pipeline to weed out any outliers with questionable ratings. The assessment validation procedure is divided into two steps: clinical-logic validation and a non-linear regression approach. The clinical-logic validation involved the development of a set of logical rules to verify the accuracy of the data. To get rid of incoherent evaluations, the locally weighted scatterplot smoothing technique was used in non-linear regression. The ML methods SVM, RF, ANN, and hybrid ANN were utilized after the 17 most important features were chosen.

Haq et al. [81] suggested a method to predict PD from speech using the SVM algorithm. This study aims to identify changes in vowel vocalization that may be used to distinguish those with PD from those who don't have PD. The MinMax Scaler and the regular scaler were used to clean up the dataset by eliminating missing values. In the feature selection step, the L1–Norm SVM method was employed to eliminate unnecessary features and increase the system's accuracy. The accuracy was highest for 10 significant characteristics. Compared to other hypermeter values, the classification performance of the SVM kernel RBF with 10 folds CV on the full features set and hyper-parameter values of $C = 1$ and $\gamma = 0.025$ was superior.

Using ML approaches, Kostov et al. [82] proposed a framework for predicting the risk of stroke disease. The purpose of this study was to use ML techniques to assess the factors for ischemic stroke in patients with epilepsy from a massive volume of data from general practitioners in Germany. Stroke-prone subpopulations were selected using the Sub-Population Optimization and Modeling Solutions (SOMS) application. To evaluate model performance, ROC was applied. Although age was not acknowledged as a significant indicator, male gender was found to be 1.5% more important than random chance.

Heart disease or cardiovascular diseases (CVD)

One of the essential parts of the body is the heart. The heart is responsible for circulating blood throughout the body [89]. The circulatory system is critical because it carries blood, oxygen, and other substances to the body's cells and tissues. Severe health conditions, including death, will result if the heart is not functioning correctly [90]. According to estimations, 17.9 million people die from CVD every year, making it the world's most prominent cause of mortality. Coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other illnesses are among the categories of heart and blood vessel disorders known as CVDs. More than 80% of all CVD deaths result from strokes and heart attacks, and 30% of these deaths occur in those younger than 70 [91, 92]. There are many types of CVDs, such as coronary heart disease, rheumatic heart disease, cerebrovascular disease, and other conditions. A critical method of lowering this toll is early identification of CVD. Using various ML approaches and data mining techniques is one of the numerous ways to improve this ailment identification and diagnosis [93]. Early identification makes it feasible to lower severe health conditions, costs, and CVD death rates. So, the purpose of this study is to conduct a comprehensive analysis of previous research concerning the prediction model for heart disease. This will be achieved by presenting details pertaining to the dataset, research objective, algorithms employed, findings, and other significant facets of the respective studies (Table 4).

In order to detect cardiac disease at an early phase, Ali et al. [94] employed six different ML algorithms on a publicly accessible UCI dataset that was gathered from Kaggle. Among 1025 instances, 51.32% of which were heart disease patients and 48.68% of which were healthy individuals. To identify outlier and extreme values during the preprocessing step, another filter known as the interquartile range (IQR) was used after substituting missing values. To eliminate outliers, the dataset was divided into three parts. After preprocessing, the accuracy of MLP, KNN, RF, DT, LR, and AdaboostM1 (ABM1) algorithms was compared. Different statistical measures were employed to assess the effectiveness of various algorithms. KNN, DT, and RF algorithms offer incredibly high accuracy.

An XGBoost-based prediction method was suggested by Shi et al. [95] to accurately detect malnutrition in children one year following congenital heart surgery. The GWC Medical Center in China provided the data, which included 536 occurrences with 15 distinct features. The continuous variables were analyzed and expressed using means and standard deviations, medians, and IQR was assessed using an independent-sample t-test or a Mann–Whitney U test. The categorical variables in this study were compared using a chi-square test and are reported as numbers and percentages. Extreme gradient boosting (XGBoost), LR, SVM, ADA, MLP, and other supervised ML methods were used. Here, the Shapley Additive exPlanations (SHAP) approach is utilized to track how each characteristic affects the outcomes of the prediction process as it is applied to each sample. The most accurate of those five algorithms was XGBoost.

Ahmed et al. [96] aimed to forecast cardiac disease based on patients' tweets using ML and big data. The primary goal of this research is to create a real-time platform that can assess and extract knowledge about heart diseases from a user's streaming tweets in order to forecast whether the person is at risk for heart disease or not. The three critical parts of the proposed system's architecture are Building an Offline Model, Stream Processing Pipeline, and Online Prediction. In the preprocessing stage, the data were scaled using the MinMax Scaler. To choose the most crucial feature subset from the data set, two feature selection techniques, Univariate feature selection, and Relief feature selection were applied. The model was trained using four classification algorithms: DT, SVM, RF, and LR, with RF providing the greatest accuracy.

Haq et al. [97] worked to develop an ML-based decision support system for the diagnosis of cardiac disease. The CHDD was employed for the forecasting model. MinMax

Table 4 Overview of different parameters from the previous works on heart disease

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	Validation	ML algorithms	Findings
Ali et al. [94]	2021	Predict heart disease	UCI dataset	1025	14	CPT, CA, age, OPK, and Thalach	Tenfold	MLP, KNN, RF, DT, LR and AB	100% accuracy using KNN, DT, and RF
Shi et al. [95]	2021	Predict postoperative malnutrition in children with heart disease	Privately collected from GWC hospital	722	15	Postoperative WAZ, discharge WAZ, preoperative WAZ, formula intake, and HIS	Fivefold	LR, SVM, ADA, MLP, and XGBoost	81% accuracy using XGBoost
Ahmed et al [96]	2019	Predict heart disease from social media posts of the patients	CHDD	303	13	THA, EIA, CPT, VCA, PES, MHR, and OPK	Tenfold	DT, SVM, RF, and LR	94.9% accuracy using RF
Haq et al. [97]	2018	Predict heart disease using hybrid intelligent system	CHDD	303	13	THA, EIA, CPT, VCA, PES, and MHR,	Tenfold	SVM, LR, KNN, ANN, NB, DT, and RF	89% accuracy using LR
Ghosh et al. [98]	2021	Predict heart disease with feature selection techniques	UCI ML repository	1190	14	Age, Trestbps, FBS, CPT, and restecg	N/A	AB, DT, GB, KNN,RFBM, DTBM, KNNBM, ABBM, GBBM, and RF	99% using RFBM
Bouk-hatem et al. [99]	2022	Predict heart disease	Kaggle	303	13	N/A	N/A	SVM, NB, MLP, and RF	The highest accuracy of 91.67%, using SVM
Mohan et al. [100]	2019	Predict heart disease using hybrid system	CHDD	303	13	N/A	N/A	DT, SVM, LR, NN, KNN, NB, HRFLM, and RF	88.4% accuracy using HRFLM
Bhatt et al. [101]	2023	Predict heart disease	Kaggle	70,000	11	N/A	Tenfold	RF, DT, MLP, and XGBoost	The highest accuracy 87.05%
Sarra et al. [102]	2022	Predict heart disease with X^2 statistical feature selection model	Two public datasets	573	14	Chest pain, exercise-induced angina, maximum heart rate, thallium, no. of major vessels, ST depression	N/A	SVM	Accuracy is 89.7%
Kadhim et al. [103]	2023	Predict heart disease	Five public datasets	1190	10	N/A	N/A	KNN, SVM, RF, and DT	The highest accuracy is 95.4% using RF
Amin et al. [104]	2018	Predict heart disease	CHDD	303	13	Sex, CP, FBS, restecg, Exang, OPK, Slope, CA, and thal	Tenfold	k-NN, DT, NB, LR, SVM, Vote, and NN	87.2% using SVM

ABBM, AdaBoost Bagging method; ADA, adaptive boosting; BPNN, back propagation neural network; CA, number of major vessels (0–3) colored by fluoroscopy; CP, chest pain; CPT, Type of chest pain; DTBM, Decision Tree Bagging method; EIA, Exercise-induced angina; Exang, Exercise induced angina (1 for yes and 0 for no); FBS, fasting blood sugar; GWC, Guangzhou Women and Children's; HIS, hospital length of stay; hr_la, Heartbeat number; HRFLM, Hybrid Random Forest with Linear Model; IHD, Ischemic heart disease; KNNBM, K-nearest neighbor Bagging method; MHR, Maximum heart rate achieved; N/A, missing or not available; OPK, ST depression induced by exercise relative to rest; PES, Slope of the peak exercise ST segment; restecg, resting electrocardiographic results; RFBM, Random Forest Bagging method; Slope, The slope of the peak exercise ST segment; THA, Thallium scan; Thalach, maximum heart rate achieved; Trestbps, resting blood pressure; VCA, Number of vessels which colored by uoroscopy; WAZ, weight for age and standard scalar were utilized in the pre-processing stage to depict ML algorithms effectively. Relief Feature Selection Algorithm, mRMR, and Least Absolute Shrinkage and Selection (LASSO) operator were the three feature selection techniques employed in this study. After choosing crucial features, seven different ML algorithms, such as SVM, LR, KNN, ANN, NB, DT, and RF, were used. RF has the highest accuracy of all of them.

A technique for effectively identifying cardiac ailment was proposed by Ghosh et al. [98]. Five separate datasets from Cleveland, Switzerland, Hungary, Statlog, and VA Long Beach are integrated into this work to create a larger, more dependable dataset for improved prediction from the UCI ML repository. Two feature selection techniques, LASSO and Relief were employed to choose the most crucial features. Five distinct algorithms were used: DT, KNN, RF, AB, and GB. To improve the system's accuracy, the authors applied ensemble techniques, including bagging and boosting. Bagging is used to lower the variance of Decision Tree classifiers. The Gradient Boost Boosting Method (GBBM) is used in this model to get the best level of accuracy.

Diabetes disease

Diabetes, characterized by a repetitive increase in blood sugar levels, has been one of the deadliest severe metabolic conditions [105, 106]. Diabetes mellitus is a collection of metabolic illnesses defined by hyperglycemia caused by abnormalities in insulin production, insulin action, or both. [107]. As many as 422 million people worldwide have diabetes, with the majority residing in poor and medium-income nations [108]. The incidence and severity of diabetes have significantly increased over the past several decades [109]. Statistics show that approximately 38.4 million people are suffering from type 2 diabetes. Among them, 29.7 million people are diagnosed, and 8.7 million are undiagnosed. On the other hand, 124.8 million people have prediabetes. Women suffer from gestational diabetes at the time of their pregnancy period. And more than 50% of them have a chance to convert this into type 2 diabetes [110]. Between 2000 and 2019, WHO found a 3% increase in diabetes mortality. However, diabetes vulnerability can be decreased by following a healthy diet and lifestyle [111]. A better quality of life and a longer lifespan are just two of the many benefits that might result from a diabetes diagnosis at an early stage [109, 112, 113]. Many researchers have made significant progress in making a proper PM for the early detection of diabetes. Therefore, this research tried to contribute to the prediction of diabetes by conducting a comprehensive study about chronic diseases, where diabetes is one of the most common chronic diseases. In this study, the PM of diabetes was analyzed from previous studies, the primary purpose of which was to find the object, dataset information, features, model validation, the algorithm used, and various other important aspects of the study (Table 5).

Hasan et al. [114] suggested a model for predicting diabetes using seven distinct ML approaches. A freely available dataset, the Pima Indian Diabetes dataset (PIDD), was utilized for this study. Mean values were utilized to replace missing values throughout the preparation stages. To get the optimal MLP design, eight distinct MLP models, ranging from one to eight hidden layers, were developed and evaluated, with the number of neurons serving as the hyperparameter for determining the best numbers. The optimal architecture was found by the MLP layout, which has 3 hidden layers (H1, H2, and H3)

Table 5 Overview of different parameters from the previous works on diabetes disease

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	Validation	ML algorithms	Findings
Hasan et al. [114]	2020	Predict diabetes	PIDD	768	8	Age, BMI, insulin, glucose, pregnancy, and Triceps	Fivefold	MLP, KNN, DT, RF, AB, NB, and XGBoost	The highest AUC is 95% using AB + XGBoost
Kopitar et al. [115]	2020	Predict type-2 diabetes	10 different Slovenian healthcare centers	3723	58	Hyper-glycemia, HDL, Tri-glycerides	N/A	LR, RF, XGBoost, Glmnet, and LightGBM	Highest RMSE 0.881 using XGBoost
Zou et al. [116]	2018	Predict diabetes mellitus	Hospital Physical Examination Data in Luzhou	68,994	14	HDL, FG, breathe, height, and LDL	Fivefold	RF, DT, and NN	Highest accuracy of 80.84% using RF
Elhadd et al. [117]	2020	Predict the glucose variability risk in patients with type 2 diabetes who fast during Ramadan	Hamad Medical Corporation	19,540	55	BMI, Ramadan, Hour Of-Day, Age, gender, and HbA1c	Fivefold	LR, RF, XGBoost, SVM, and DL	The highest R^2 value is 0.836 by XGBoost
Tigga et al. [118]	2020	Predict type-2 diabetes	Privately collected by survey	952	18	Age, family diabetes, physical activity, Regular Medicine, and gestation diabetes	Tenfold	LR, KNN, SVM, DT, NB, and RF	Highest accuracy94.1% using RF
Islam et al. [119]	2023	Predict diabetes and Provide clinical decision	PIDD	768	8	Glucose, BMI, insulin, BP, age	Tenfold	KNN, RF, SVM, NB, HBGB, and DT	Highest accuracy 92.21% using HBGB
Krishna-mouth et al. [120]	2022	Employing big data and ML for diabetes prediction	PIDD	768	8	BMI, age, and glucose level	N/A	LR, KNN DT, RF, SVM	Highest accuracy 83% using RF and SVM
Reddy et al. [121]	2020	Predict diabetes	PIDD	768	8	N/A	Tenfold	LR, SVM, KNN, RF, GB and NB	Highest accuracy 98.48% from RF
Dagliati et al. [122]	2018	Predict diabetes patient complications	ICSM hospital	943	N/A	BMI, hypertension	N/A	LR, NB, SVM, and RF	Highest accuracy77.7% of 3 years retinopathy
Laila et al. [123]	2022	Predict Early-Stage Diabetes	UCI repository	520	17	N/A	Tenfold	AdaBoost, Bagging, RF	Highest accuracy 97.11% using RF

Table 5 (continued)

Author	Year	Objective	Dataset source	Dataset instance	Dataset features	Important features	Validation	ML algorithms	Findings
Wee et al. [124]	2023	Predict diabetes	PIDD	768	8	N/A	Tenfold	SVM, LR, RF, CNN, DNN, and MLP	The highest accuracy 98.1%
Wang et al. [125]	2020	Predict hypoglycemic drugs for type 2 diabetes patients	Chinese PLA General Hospital	2443	17	ALT, AST, TP, SA, urea, Glutamyltransferase, Creatinine, Glucose, Triglyceride, SUA, Creatine kinase, Potassium, and AP	N/A	SVM	Average precision 86.95%
Pattnayak et al. [126]	2024	Predict diabetes disease	PIDD	768	8	N/A	N/A	NB, KNN, and LR	The highest precision was 95% using LR

ALT, Alanine Aminotransferase; AP, Alkaline Phosphatase; dlab_pred, diabetes pedigree function; FG, fasting blood glucose; Glmnet, generalized linear model; HBGB, histogram-based gradient boosting, HbA1c, Hemoglobin A1C; ICSM, Istituto Clinico Scientifico Maugeri; LightGBM, Light Gradient Boosting Machine; N/A, missing or not available; SA, Serum Albumin; TP, Total Protein; Triceps, Triceps Skin Fold Thickness (mm) with 16,64 and 64 neurons, respectively, and was selected using a grid search approach. Compared to the other six ML approaches, XGBoost + AB provided better accuracy.

The goal of the type 2 diabetes prediction model by Kopitar et al. [115] was to use various ML methods to identify individuals at an early stage. The primary purpose of this research was to determine if ML-based methods could be used to accurately forecast impaired fasting glucose and fasting plasma glucose level values at an early stage. Data for this study was gathered from ten different Slovenian medical facilities. In the preprocessing stage, the number of samples is reduced by maintaining all cases with significant absolute values of gradients and arbitrarily picking examples with smaller absolute values of gradients. The system was evaluated using both AUC and AUPRC since the dataset was unbalanced.

In [116], Zou et al. introduced a methodology for diabetes prediction based on three distinct ML approaches. The purpose of this research was to forecast diabetes and make a comparison of their private dataset to the PIDD, as well as to identify the important factors that determine prediction system accuracy for both datasets. PCA and mRMR are used to select important features. From the three ML algorithms, WEKA implemented DT and RF, while MATLAB implemented NN. Five features-height, breathing, FG, high-density lipoprotein (HDL), and low-density lipoprotein (LDL) were chosen to test accuracy based on the results of the mRMR technique.

Elhadd et al. [117] employed several ML approaches to predict metabolic outcomes with type 2 diabetes (T2DM) who fasted throughout Ramadan. T2DM patients with hypoglycemia were recruited before and throughout Ramadan and used ML approaches that integrate data on glucose variability using the glucose monitoring system and physical activity using the Fitbit-flex 2. The SHAP plot was utilized in this case to classify the relevance of various characteristics. The model's performance was calculated by mean absolute error (MAE), where data points closer to 0 were ideal, and coefficient of determination (R^2), where values near 1 were ideal. ML model XGBoost had the best performance. While 0.837 and 17.47 were the R^2 and MAE values, respectively.

Another method for predicting diabetes was suggested by Tigga and Garg [118], where authors used six different ML algorithms (LR, KNN, SVM, DT, NB, and RF). The authors used a self-prepared questionnaire to collect data through a survey. There were 952 participants in that survey. Because the dataset contains more variables pertinent to determining the risk of diabetes, the accuracy of the authors' model, as measured by comparison with PIDD, is most significant. RF algorithm was the most accurate compared to the other five ML techniques used in their study.

Islam et al. [119] proposed a decision support system for diabetes patients after predicting diabetes by using six ML classifiers. After preprocessing PIDD, the authors used rule-based approaches, which helped them to achieve 92.21% accuracy with a histogram-based GB algorithm. CDSS was implemented where users can give the required input parameters through a web-based user interface to get decision support if the patient has diabetes or a comparative graph of some critical parameters based on essential features such as BMI, BP, insulin level, glucose, and skin thickness for the non-diabetic patient.

Publicly available datasets for CD classification

Table 6 provides the list of datasets publicly available for the classification of different chronic diseases such as cancer, liver disease, brain disease, heart disease and diabetes disease from this study.

Discussion

This comprehensive research found that ML algorithms have shown promising outcomes in the prediction of CD. It can be seen that a significant portion of the studies used public datasets. Several researchers did outstanding work and achieved the highest accuracy. However, compared to authors who used public datasets, the authors who had

Table 6 Publicly available dataset for different chronic disease classifications

Disease name	Dataset name	Source	Instances	No. of features	Remark
Cancer	Ovarian cancer dataset	Kaggle	349	51	All the patients in this dataset were diagnosed by pathology after surgery. None of the patients received any type of pre-operative
	Skin cancer dataset	Kaggle	2357	N/A	All images were sorted according to the classification taken with ISIC
Liver Disease	Hepatitis C dataset	UCI repository	615	12	The dataset contains patients' laboratory values of blood donors and demographic values
	Indian liver patient dataset	UCI repository	583	10	The dataset's patient records collected from the India
Brain disease	Alzheimer disease dataset	Kaggle	5000	N/A	All the images are MRI images collected from various sources
	Parkinson disease dataset	UCI repository	5875	19	This dataset is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month
Heart disease	Heart disease dataset	Kaggle	70,000	11	There are 3 types of input features: factual information, results of medical examination, patient's information
	Cleveland heart disease dataset	UCI repository	303	14	In this dataset 8 values are nominal and 5 values are numeric
Diabetes disease	Pima Indian diabetes dataset	UCI repository	768	8	All the samples are female whose age is at least 21 years old
	Early-stage diabetes risk prediction dataset	UCI repository	520	17	This dataset comprises sign and symptom information of patients who are newly diagnosed with diabetes or who are at risk of developing the condition

access to private datasets showed greater accuracy, as well as had an improved result in other performance matrix values. The biggest disadvantage of using a publicly available dataset is the minimal quantity of data samples. Having a sufficiently big training dataset is a fundamental prerequisite when employing classification algorithms to simulate a disease. In order to validate the estimators reasonably, an equitable-sized dataset must be split into training and testing sets. An unbalanced dataset, numerous missing values, and the existence of outliers are other factors that reduce the accuracy of a publicly available dataset. Enhanced accuracy can be achieved by proper preprocessing of the dataset. Most of the authors exerted considerable effort in the preprocessing phase, which included deleting missing values, scaling the dataset, balancing the data, and removing outliers, which were able to increase accuracy. Data scalability led to improved convergence, which allowed authors to achieve an accuracy of 95% or better. SMOTE and random over-sampling were the two most prevalent strategies for balancing datasets in our scrutinized research papers. It is essential to the model-building process to narrow down the features to a manageable number. Almost every author employed a variety of strategies to identify significant features. However, to get a more exact and accurate subset of characteristics, a few authors used several feature selection methods to get a smaller subset of features that was more precise and relevant to their study. The most widely used strategies for selecting features in our reviewed papers are LASSO, RELIEF, and mRMR. Almost all of the studies mentioned here conducted validation tests to evaluate the efficacy of their learning algorithms. A significant factor in the encouraging outcomes of multiple experiments was the employment of various ML approaches with the intention of identifying the most effective one. A customized ensemble approach increased the accuracy of some authors. This study found that SVM and RF classifiers were two of the most popular ML algorithms for predicting cancer patient outcomes. Numerous studies have shown that SVM, DT, and NB are superior to other methods for predicting CVD. For the purpose of predicting liver disease, several boosting algorithms were mainly employed. Again, SVM and RF were widely used for predicting brain disease and diabetes as well.

Conclusion

The recent research on ML-based techniques for predicting CD was focused on in this article. Among them, certain authors have accomplished remarkable feats. As summarized in Tables 1, 2, 3, 4, 5, it can be seen that various important characteristics of PM using ML techniques like no. of features, dataset information, validation technique used, important features, as well as the objective and the findings which various researchers have found on different CDs such as liver disease, cancer, brain disease, heart disease, and diabetes disease have been investigated respectively. From these tabular data, a researcher can get a precise overview of the previous work of PM on CD as well as the diagnosis outcome discussed by the previous researchers. Additionally, this study also represents the list of available datasets that the researchers can work with for further research. Thus, it will definitely improve the work speed, and it can bring new ideas for the betterment of the healthcare domain. This study also finds that most suggested research in the last several years has been on creating PM for CD through the use of supervised ML techniques and classification algorithms. We have found from previous studies that SVM and RF classifiers were the most popular ML algorithms for predicting CD. Despite all of the findings of this study, there are some confounding factors as well. The main limitation of this research is that the study focuses only on the prediction model, which ultimately limits the scope of the in-depth insight into a particular chronic disease. Furthermore, our search strategy could prevent this study from covering a broad range of results regarding the previous study, so different search titles may give different results. However, it is recommended that all of the limitations be addressed, further study about the broad range of areas, and an in-depth analysis of the studies of chronic diseases be conducted. From this research, we also recommend that the future development of the prediction models to create a proper CDSS for remotely monitoring the patients of CD would be particularly advantageous for both patients and physicians because previous studies suggest that these patients need to be observed on a frequent basis. Better results and outcomes, as well as effective patient treatment for CD, will result from the PM's seamless integration with hospitals and medical domains to provide consistent health records and data. It has the potential to not only enhance the existing healthcare system but also to make medical care more accessible to everyone by lowering the costs associated with providing treatment.

Abbreviations

ML	Machine learning
CD	Chronic diseases
PM	Prediction model
HI	Healthcare industry
WHO	World Health Organization
CVD	Cardiovascular disease
ALFSG	Acute liver failure study group
IQR	Interquartile range
PIDD	Pima Indian diabetes dataset
HDL	High-density lipoprotein
LDL	Low-density lipoprotein
T2DM	Type 2 diabetes
MAE	Mean absolute error
CHDD	Clevenand heart disease dataset
CDSS	Clinical decision support systems
SOMS	Sub-population optimization and modeling solutions
AUPRC	Area under precision-recall curve
LASSO	Least absolute shrinkage and selection
mRMR	Minimal-redundancy-maximum-relevance
SVM	Support vector machine
RF	Random forest
DT	Decision tree
NB	Naïve Bayes
LR	Linear regression
KNN	K-nearest neighbor
DL	Deep learning
ANN	Artificial neural network
MLP	Multilayer perceptron
LSTM	Long short-term memory networks
SGD	Stochastic gradient descent
MLPNN	Multilayer perceptron neural network
CHAID	Chi-square automatic interaction detector
ROC	Receiver operating curve
GB	Gradient boosting
PNN	Probabilistic neural networks
GBBM	Gradient boost boosting method

Acknowledgements

Not applicable.

Author contributions
RI and AS wrote the manuscript; MRI participated in the improvement and finalizing of the manuscript. All authors read and approved the final manuscript.

Funding
The authors received no financial support for the research authorship and publication of this article.

Availability of data and materials
The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Competing interests
The authors declare that they have no conflicts of interest.

Received: 7 November 2023 Accepted: 27 June 2024
Published online: 16 July 2024

References

1. Davenport T, Kalakota R (2019) DIGITAL TECHNOLOGY The potential for artificial intelligence in healthcare
2. Bekkers E n.d. Machine learning 1 Lecture 1.2-What is Machine Learning?
3. Horvitz E, Mulligan D (2015) Data, privacy, and the greater good. Science 349(6245):253–255. https://doi.org/10.1126/science.aac4520
4. Allenbrand C (2024) Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews. Healthc Anal 5:100288. https://doi.org/10.1016/j.health.2023.100288
5. Bi Q, Goodman KE, Kaminsky J, Lessler J (2019) What is machine learning? A primer for the epidemiologist. Am J Epidemiol. https://doi.org/10.1093/aje/kwz189
6. Devi MK et al (2022) Design and implementation of advanced machine learning management and its impact on better healthcare services: a multiple regression analysis approach (MRAA). Comput Math Methods Med. https://doi.org/10.1155/2022/2489116
7. Adlung L, Cohen Y, Mor U, Elinav E (2021) Machine learning in clinical decision making. Medicine 2(6):642–665. https://doi.org/10.1016/j.medj.2021.04.006
8. Binson VA, Thomas S, Subramoniam M, Arun J, Naveen S, Madhu S (2024) A review of machine learning algorithms for biomedical applications. Ann Biomed Eng 52(5):1159–1183. https://doi.org/10.1007/s10439-024-03459-3
9. Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J (2017) Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol 48(3):128–138. https://doi.org/10.14445/22312803/IJCTT-V48P126
10. Sun L, Gupta RK, Sharma A (2022) Review and potential for artificial intelligence in healthcare. Int J Syst Assur Eng Manag 13(S1):54–62. https://doi.org/10.1007/s13198-021-01221-9
11. Yach D, Hawkes C, Linn Gould C, Hofman KJ The Global burden of chronic diseases overcoming impediments to prevention and control. [Online]. Available: http://jama.jamanetwork.com/
12. Bernell S, Howard SW (2016) Use Your words carefully: What is a chronic disease? Front Public Health. https://doi.org/10.3389/fpubh.2016.00159
13. Yan Y, Mi J (2021) Noncommunicable chronic disease prevention should start from childhood. Pediatr Invest 5(1):3–5. https://doi.org/10.1002/ped4.12254
14. Durstine JL, Gordon B, Wang Z, Luo X (2013) Chronic disease and the link to physical activity. J Sport Health Sci 2(1):3–11. https://doi.org/10.1016/j.jshs.2012.07.009
15. Noncommunicable diseases. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases Accessed 11 Jul 2023
16. World Health Organization (2005) WHO steps surveillance manual : the WHO stepwise approach to chronic disease risk factor surveillance. WHO
17. Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: a review. Egypt. Inform J 19(3):179–189. https://doi.org/10.1016/j.eij.2018.03.002
18. Nusinovici S et al (2020) Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol 122:56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002
19. Kim C, Son Y, Youm S (2019) Chronic disease prediction using character-recurrent neural network in the presence of missing information. Appl Sci (Switzerland). https://doi.org/10.3390/app9102170
20. Ketkar Y, Gawade S (2022) A decision support system for selecting the most suitable machine learning in healthcare using user parameters and requirements. Healthc Anal 2:100117. https://doi.org/10.1016/j.health.2022.100117
21. Jahandideh S, Ozavci G, Sahle BW, Kouzani AZ, Magrabi F, Bucknall T (2023) Evaluation of machine learning-based models for prediction of clinical deterioration: a systematic literature review. Int J Med Inform 175:105084. https://doi.org/10.1016/j.ijmedinf.2023.105084
22. Kumar NK, Sikamani KT (2020) Prediction of chronic and infectious diseases using machine learning classifiers—a systematic approach. Int J Intell Eng Syst 13(4):11–20. https://doi.org/10.22266/IJIES2020.0831.02
23. Usman SM, Usman M, Fong S (2017) Epileptic seizures prediction using machine learning methods. Comput Math Methods Med. https://doi.org/10.1155/2017/9074759

24. Simon S, Kolyada N, Akiki C, Potthast M, Stein B, Siegmund N (2023) Exploring hyperparameter usage and tuning in machine learning research. In: 2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN), May 2023, Published, https://doi.org/10.1109/cain58948.2023.00016
25. Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. https://doi.org/10.1007/s42979-021-00592-x
26. Lin SS, Shen SL, Zhou A, Xu YS (2021) Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. Autom Construct. https://doi.org/10.1016/j.autcon.2020.103490
27. Herhausen D, Bernritter SF, Ngai EWT, Kumar A, Delen D (2024) Machine learning in marketing: recent progress and future research directions. J Bus Res 170:114254. https://doi.org/10.1016/j.jbusres.2023.114254
28. Hofhuis J, Gonçalves J, Schafraad P, Wu B (2024) Examining strategic diversity communication on social media using supervised machine learning: development, validation and future research directions. Public Relat Rev 50(1):102431. https://doi.org/10.1016/j.pubrev.2024.102431
29. Liu M, Xue J, Zhao N, Wang X, Jiao D, Zhu T (2021) Using social media to explore the consequences of domestic violence on mental health. J Interpers Violence 36(3–4):1965–1985. https://doi.org/10.1177/0886260518757756
30. Türkbayrağí MG, Dogu E, Esra Albayrak Y (2022) Artificial intelligence based prediction models: sales forecasting application in automotive aftermarket. J Intell Fuzzy Syst 42(1):213–225. https://doi.org/10.3233/JIFS-219187
31. MacKay C, Klement W, Vanberkel P, Lamond N, Urquhart R, Rigby M (2023) A framework for implementing machine learning in healthcare based on the concepts of preconditions and postconditions. Healthc Anal 3:100155. https://doi.org/10.1016/j.health.2023.100155
32. Allgaier J, Mulansky L, Draelos RL, Pryss R (2023) How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med 143:102616. https://doi.org/10.1016/j.artmed.2023.102616
33. Abdel-Misih SRZ, Bloomston M (2010) Liver anatomy. Surg Clin North Am 90(4):643–653. https://doi.org/10.1016/j.suc.2010.04.017
34. Asrani SK, Devarbhavi H, Eaton J, Kamath PS (2019) Burden of liver diseases in the world. J Hepatol 70(1):151–171. https://doi.org/10.1016/j.jhep.2018.09.014
35. Mokdad AA et al (2014) Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med. https://doi.org/10.1186/s12916-014-0145-y
36. Zeng DY et al (2021) Global burden of acute viral hepatitis and its association with socioeconomic development status, 1990–2019. J Hepatol 75(3):547–556. https://doi.org/10.1016/j.jhep.2021.04.035
37. Hepatitis B (2023) https://www.who.int/news-room/fact-sheets/detail/hepatitis-b Accessed 19 Jul 2023
38. Rinella ME (2015) Nonalcoholic fatty liver disease a systematic review. JAMA—J Am Med Assoc 313(22):2263–2273. https://doi.org/10.1001/jama.2015.5370
39. Liu YX et al (2021) Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: an extended study. Hepatobiliary Pancreat Dis Int 20(5):409–415. https://doi.org/10.1016/j.hbpd.2021.08.004
40. Liu R et al (2023) Prediction model for hepatocellular carcinoma recurrence after hepatectomy: machine learning-based development and interpretation study. Heliyon 9(11):e22458. https://doi.org/10.1016/j.heliyon.2023.e22458
41. Hashem S et al (2018) Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients. IEEE/ACM Trans Comput Biol Bioinform 15(3):861–868. https://doi.org/10.1109/TCBB.2017.2690848
42. Harrison E, Chang M, Hao Y, Flower A (2018) Using machine learning to predict near-term mortality in cirrhosis patients hospitalized at the University of Virginia health system. In: 2018 Systems and Information Engineering Design Symposium (SIEDS), Apr. 2018, https://doi.org/10.1109/sieds.2018.8374719
43. Speiser JL, Karvellas CJ, Wolf BJ, Chung D, Koch DG, Durkalski VL (2019) Predicting daily outcomes in acetaminophen-induced acute liver failure patients with machine learning techniques. Comput Methods Programs Biomed 175:111–120. https://doi.org/10.1016/j.cmpb.2019.04.012
44. Safdari R, Deghatipour A, Gholamzadeh M, Maghooli K (2022) Applying data mining techniques to classify patients with suspected hepatitis C virus infection. Intell Med 4:193–198. https://doi.org/10.1016/j.imed.2021.12.003
45. Goldar SZ, Rikhtegar Ghiasi A, Badamchizadeh MA, Khoshbaten M (2020) An ANFIS-PSO algorithm for predicting four grades of non-alcoholic fatty liver disease. In: 2020 International congress on human-computer interaction, optimization and robotic applications (HORA), Jun. 2020, https://doi.org/10.1109/hora49412.2020.9152881
46. Barus OP, Happy J, Jusin JJ, Pangaribuan SZ, Nadjar HF (2022) Liver disease prediction using support vector machine and logistic regression model with combination of PCA and SMOTE. In: 2022 1st International conference on technology innovation and its applications (ICTIIA), Tangerang, Indonesia, 2022, pp. 1–6, https://doi.org/10.1109/ICTIIA54654.2022.9935879
47. Feldman TC, Dienstag JL, Mandl KD, Tseng YJ (2021) Machine-learning-based predictions of direct-acting antiviral therapy duration for patients with hepatitis C. Int J Med Inform. https://doi.org/10.1016/j.ijmedinf.2021.104562
48. Gupta K, Jiwani N, Afreen N, Divyarani D (2022) Liver disease prediction using machine learning classification techniques. In: 2022 IEEE 11th International conference on communication systems and network technologies (CSNT), Apr. 2022, https://doi.org/10.1109/csnt54456.2022.9787574
49. Ahad AA, Das B, Khan MR, Saha N, Zahid A, Ahmad M (2024) Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling. Results Eng 22:102059. https://doi.org/10.1016/j.rineng.2024.102059
50. Goel S et al (2011) Normalization of the vasculature for treatment of cancer and other diseases. Physiol Rev 91:1071–1121. https://doi.org/10.1152/physrev.00038.2010.-New
51. Seyfried TN, Shelton LM (2010) Cancer as a metabolic disease. [Online]. Available: http://www.nutritionandmetabolism.com/content/7/1/7
52. Gonzalez H, Hagerling C, Werb Z (2018) Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev 19–20:1267–1284. https://doi.org/10.1101/gad.314617.118

53. Ravasco P, Monteiro-Grillo I, Marques Vidal P, Camilo ME (2004) Cancer: disease and nutrition are key determinants of patients' quality of life. Support Care Cancer 12(4):246–252. https://doi.org/10.1007/s00520-003-0568-z
54. Sung H et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin 71(3):209–249. https://doi.org/10.3322/caac.21660
55. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005
56. Cox TR, Erler JT (2011) Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. DMM Dis Models Mech 4(2):165–178. https://doi.org/10.1242/dmm.004077
57. Abbasi EY et al (2023) Optimizing skin cancer survival prediction with ensemble techniques. Bioengineering 11(1):43. https://doi.org/10.3390/bioengineering11010043
58. JM SL, Subbulakshmi P (2024) Innovative approach towards early prediction of ovarian cancer: machine learning—enabled Xai Techniques. Heliyon [Preprint]. https://doi.org/10.1016/j.heliyon.2024.e29197
59. Murugan A, Nair SAH, Preethi AAP, Kumar KPS (2021) Diagnosis of skin cancer using machine learning techniques. Microprocess Microsyst. https://doi.org/10.1016/j.micpro.2020.103727
60. Naji MA, el Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O (2021) Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Comput Sci 191:487–492. https://doi.org/10.1016/j.procs.2021.07.062
61. Sakai Y et al (2018) Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network. Annu Int Conf IEEE Eng Med Biol Soc. https://doi.org/10.1109/EMBC.2018.8513274
62. Salmi N, Rustam Z (2019) Naïve Bayes classifier models for predicting the colon cancer. IOP Conf Ser Mater Sci Eng. https://doi.org/10.1088/1757-899X/546/5/052068
63. Hasan I, Ali S, Rahman H, Islam K (2022) Automated detection and characterization of colon cancer with deep convolutional neural networks. J Healthc Eng 2022:1–12. https://doi.org/10.1155/2022/5269913
64. Adeoye J, Hui L, Koohi-Moghadam M, Tan JY, Choi SW, Thomson P (2022) Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis. Int J Med Inform. https://doi.org/10.1016/j.ijmedinf.2021.104635
65. Xie Y et al (2021) Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl Oncol. https://doi.org/10.1016/j.tranon.2020.100907
66. Gupta P et al (2019) Prediction of colon cancer stages and survival period with machine learning approach. Cancers (Basel). https://doi.org/10.3390/cancers11122007
67. Mourad M et al (2020) Machine Learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis. Sci Rep. https://doi.org/10.1038/s41598-020-62023-w
68. Mayfield brain & spine (2023) Mayfieldclinic.com. https://mayfieldclinic.com/pe-anatbrain.htm Accessed 03 Aug 2023
69. Stiles J, Jernigan TL (2010) The basics of brain development. Neuropsychol Rev 20(4):327–348. https://doi.org/10.1007/s11065-010-9148-4
70. Mattson MP, Duan W, Pedersen WA, Culmsee C (2001) Neurodegenerative disorders and ischemic brain diseases
71. Vanlandewijck M et al (2018) A molecular atlas of cell types and zonation in the brain vasculature. Nature 554(7693):475–480. https://doi.org/10.1038/nature25739
72. Sosin DM (1995) Trends in death associated with traumatic brain injury 1979 through 1992. JAMA. https://doi.org/10.1001/jama.1995.03520460060036
73. Cummings JL (2002) Alzheimer disease. JAMA. https://doi.org/10.1001/jama.287.18.2335
74. Poewe W et al (2017) Parkinson disease. Nat Rev Dis Primers 3:1–21. https://doi.org/10.1038/nrdp.2017.13
75. Shi K, Tian D-C, Li Z-G, Ducruet AF, Lawton MT, Shi F-D (2019) Global brain inflammation in stroke. Lancet Neurol 18(11):1058–1066. https://doi.org/10.1016/s1474-4422(19)30078-x
76. Johnson W, Onuma O, Owolabi M, Sachdev S (2016) Stroke: a global response is needed. Bull World Health Org 94(9):634–635. https://doi.org/10.2471/BLT.16.181636
77. Lo EH, Dalkara T, Moskowitz MA (2003) Neurological diseases: mechanisms, challenges and opportunities in stroke. Nat Rev Neurosci 4(5):399–414. https://doi.org/10.1038/nrn1106
78. Rangaswamy U, Dharshini SAP, Yesudhas D, Gromiha MM (2020) VEPAD—Predicting the effect of variants associated with Alzheimer's disease using machine learning. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2020.103933
79. Rani P et al. (2024) A machine learning model for alzheimer's disease prediction. IET Cyber-Phys Syst Theory Appl [Preprint]. https://doi.org/10.1049/cps2.12090
80. Lin CH et al (2020) Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Comput Methods Programs Biomed. https://doi.org/10.1016/j.cmpb.2020.105381
81. Haq AU et al (2019) Feature selection based on L1-norm support vector machine and effective recognition system for parkinson's disease using voice recordings. IEEE Access 7:37718–37734. https://doi.org/10.1109/ACCESS.2019.2906350
82. Kostev K, Wu T, Wang Y, Chaudhuri K, Tanislav C (2021) Predicting the risk of stroke in patients with late-onset epilepsy: a machine learning approach. Epilepsy Behav. https://doi.org/10.1016/j.yebeh.2021.108211
83. Dritsas E, Trigka M (2022) Stroke risk prediction with machine learning techniques. Sensors. https://doi.org/10.3390/s22134670
84. Shoily TI, Islam T, Jannat S, Tanna SA, Alif TM, Ema RR (2019) Detection of stroke disease using machine learning algorithms. In: 2019 10th International conference on computing, communication and networking technologies (ICCCNT), July 2019, https://doi.org/10.1109/icccnt45670.2019.8944689
85. Martinez-Murcia FJ, Ortiz A, Gorriz JM, Ramirez J, Castillo-Barnes D (2020) Studying the manifold structure of alzheimer's disease: a deep learning approach using convolutional autoencoders. IEEE J Biomed Health Inform 24(1):17–26. https://doi.org/10.1109/JBHI.2019.2914970
86. Grover S, Bhartia S, Akshama AY, Seeja KR (2018) Predicting severity of parkinson's disease using deep learning. Procedia Comput Sci 132:1788–1794. https://doi.org/10.1016/j.procs.2018.05.154

87. Sudharsan M, Thailambal G (2021) Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA). Mater Today Proc. https://doi.org/10.1016/j.matpr.2021.03.061
88. Franciotti R et al (2023) Comparison of machine learning-based approaches to predict the conversion to alzheimer's disease from mild cognitive impairment. Neuroscience 514:143–152. https://doi.org/10.1016/j.neuroscience.2023.01.029
89. Hoffman JIE, Kaplan S (2002) The incidence of congenital heart disease. J Am Coll Cardiol 39(12):1890–1900. https://doi.org/10.1016/S0735-1097(02)01886-7
90. Hazra A, Mandal SK, Gupta A, Mukherjee A, Mukherjee A (2017) Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. [Online]. Available: http://www.ripublication.com
91. Tsao CW et al (2022) Heart disease and stroke statistics-2022 update: a report from the American heart association. Circulation 145(8):153–639. https://doi.org/10.1161/CIR.0000000000001052
92. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases Accessed 26 Aug 2023
93. Khan SU et al (2022) A comparative analysis of premature heart disease—and cancer-related mortality in women in the USA, 1999–2018. Eur Heart J Qual Care Clin Outcomes 8(3):315–323. https://doi.org/10.1093/ehjqcco/qcaa099
94. Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JMW, Moni MA (2021) Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2021.104672
95. Shi H et al (2022) Explainable machine learning model for predicting the occurrence of postoperative malnutrition in children with congenital heart disease. Clin Nutr 41(1):202–210. https://doi.org/10.1016/j.clnu.2021.11.006
96. Ahmed H, Younis EMG, Hendawi A, Ali AA (2020) Heart disease identification from patients' social posts, machine learning solution on Spark. Futur Gener Comput Syst 111:714–722. https://doi.org/10.1016/j.future.2019.09.056
97. Haq AU, Li JP, Memon MH, Nazir S, Sun R, Garciá-Magarinõ I (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mob Inf Syst. https://doi.org/10.1155/2018/3860146
98. Ghosh P et al (2021) Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. IEEE Access 9:19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759
99. Boukhatem C, Youssef HY, Nassif AB (2022) Heart disease prediction using machine learning. In: 2022 advances in science and engineering technology international conferences, ASET 2022, 2022. https://doi.org/10.1109/ASET53988.2022.9734880
100. 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
101. Bhatt CM et al (2023) Effective heart disease prediction using machine learning techniques. Algorithms 16(2):88. https://doi.org/10.3390/a16020088
102. Sarra R et al (2022) Enhanced heart disease prediction based on machine learning and X2 statistical optimal feature selection model. Designs 6(5):87. https://doi.org/10.3390/designs6050087
103. Abood Kadhim M, Radhi AM (2023) Heart disease classification using optimized machine learning algorithms. Iraqi J Comput Sci Math. https://doi.org/10.52866/ijcsm.2023.02.02.004
104. Amin MS, Chiam YK, Varathan KD (2019) Identification of significant features and data mining techniques in predicting heart disease. Telemat Inform 36:82–93. https://doi.org/10.1016/j.tele.2018.11.007
105. Epstein M, Sowers JR (1992) Diabetes mellitus and hypertension. Hypertension 19(5):403–418. https://doi.org/10.1161/01.HYP.19.5.403
106. Atkinson MA, Eisenbarth GS (2001) Type 1 diabetes: new perspectives on disease pathogenesis and treatment. Lancet 358(9277):221–229. https://doi.org/10.1016/S0140-6736(01)05415-0
107. Eisenbarth GS (1986) Type I diabetes mellitus. A chronic autoimmune disease. N Engl J Med 314(21):1360–1368. https://doi.org/10.1056/NEJM198605223142106
108. Kharroubi AT (2015) Diabetes mellitus: the epidemic of the century. World J Diabet 6(6):850. https://doi.org/10.4239/wjd.v6.i6.850
109. Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 2 Sep 2023
110. Diabetes Statistics. (2024) National institute of diabetes and digestive and kidney diseases. https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics
111. Hu FB et al (1999) Prospective study of adult onset diabetes mellitus (type 2) and risk of colorectal cancer in women. J Natl Cancer Inst 91(6):542–547
112. Stumvoll M, Goldstein BJ, van Haeften TW (2005) Type 2 diabetes: principles of pathogenesis and therapy. The Lancet 9467:1333–1346. https://doi.org/10.1016/s0140-6736(05)61032-x
113. Doğru A, Buyrukoğlu S, Arı M (2023) A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Med Biol Eng Comput 61(3):785–797. https://doi.org/10.1007/s11517-022-02749-z
114. 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
115. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G (2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. https://doi.org/10.1038/s41598-020-68771-z
116. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) Predicting diabetes mellitus with machine learning techniques. Front Genet. https://doi.org/10.3389/fgene.2018.00515
117. Elhadd T et al (2020) Artificial Intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (The PROFAST-IT Ramadan study). Diabet Res Clin Pract. https://doi.org/10.1016/j.diabres.2020.108388
118. Tigga NP, Garg S (2020) Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci 167:706–716. https://doi.org/10.1016/j.procs.2020.03.336
119. Islam R, Sultana A, Tuhin MdN, Saikat MdSH, Islam MR (2023) Clinical decision support system for diabetic patients by predicting type 2 diabetes using machine learning algorithms. J Healthc Eng. https://doi.org/10.1155/2023/6992441

120. Krishnamoorthi R et al (2022) A novel diabetes healthcare disease prediction framework using machine learning techniques. J Healthc Eng. https://doi.org/10.1155/2022/1684017
121. Jashwanth Reddy D et al (2020) Predictive machine learning model for early detection and analysis of diabetes. Mater Today Proc. https://doi.org/10.1016/j.matpr.2020.09.522
122. Dagliati A et al (2018) Machine learning methods to predict diabetes complications. J Diabet Sci Technol 12(2):295–302. https://doi.org/10.1177/1932296817706375
123. Laila UE, Mahboob K, Khan AW, Khan F, Taekeun W (2022) An ensemble approach to predict early-stage diabetes risk using machine learning: an empirical study. Sensors. https://doi.org/10.3390/s22145247
124. Wee BF et al (2023) Diabetes detection based on machine learning and deep learning approaches. Multim Tools Appl 83(8):24153–24185. https://doi.org/10.1007/s11042-023-16407-5
125. Wang X, Yang Y, Xu Y, Chen Q, Wang H, Gao H (2020) Predicting hypoglycemic drugs of type 2 diabetes based on weighted rank support vector machine. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2020.105868
126. Pattnayak P, Patra SS, Patnaik S (2024) Diabetic Patient diagnosis through the use of machine learning techniques. In: 2024 5th International conference on mobile computing and sustainable informatics (ICMCSI), Jan. 2024, Published, https://doi.org/10.1109/icmcsi61536.2024.00073

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.