International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Multi Disease Prediction using Machine Learning: A Survey

P. Akshay [1], S. Swathi [2], P. Yaswanth [3], V. Anil Kumar [4], P. Krishna Vamsi [5], P. Someswari [6]

[1,2,3,4,5] Student, GMRIT Rajam, Rajam Pincode: 532127 India

[6]Assistant Professor, GMRIT Rajam, Rajam Pincode: 532127 India

ABSTRACT

The healthcare domain is one of the prominent research fields in the current scenario with the rapid improvement of technology and data. Machine Learning is an emerging approach that helps in prediction and diagnosis of a disease. With the rapid development of technology and data, the healthcare sector is currently one of the most important study areas. Methods which are proposed by the authors to assist in disease detection and prediction is machine learning. Machine learning is used to predict if a person has a disease based on symptoms within the given data set. Machine learning algorithms like Naive Bayes, Decision Tree, Random Forest, Convolution Neural Network, Support Vector Machine, and Ensemble models are used to predict the disease. For each disease, a new model was created using the specified algorithms known as the hybrid model, which is used to forecast whether or not the person is suffering from disease. Even a user interface has been given for patients so that they may manually enter their symptoms and determine whether or not they are suffering from it. This user interface covers popular web topics, most notably React, CSS, and Flask

Keywords: Machine learning, CNN, Disease prediction, Hybrid model, Flask.

1. Introduction

There are as many diseases that humans face all around the world, and their ultimate task is to overcome these difficulties and overcome them. They usually go to hospitals and visit doctors. Many people are afflicted with various ailments. Disease prediction can be used to identify patients who are at risk for illness or other health problems. The quality of care will improve and potential hospital stays may be avoided as a result of the clinicians taking the appropriate precautions to avoid or lessen the risk. The recent developments in data analytics tools and methods, disease risk prediction may now make extensive use of semantic data, including demographics, clinical diagnosis and measurements, health habits, test results, prescriptions, and service utilization. Building illness prediction models using electronic health data may be a practical solution in this area. Also, if we have any health problem we usually go to the nearest hospital, but those hospitals may not be vacant all the time so if we want to know what the disease was and who we should consult, this model provides a user interface in which the patient can enter their symptoms and learn whether they have the disease or not, as well as who they should consult. Many diseases exist across the world, yet only a few were predicted: heart disease, lung disease, vector-borne diseases, monkey pox, and malaria. As there are various diseases to predict, there will be several datasets used in this prediction, with the majority of these datasets coming from the Kaggle database. Both textual and image data are included in the datasets. There will be a hybrid model for each disease, and this hybrid model will be built using appropriate ensemble learning methods and different algorithms such as SVM, KNN, NB, RF and some deep learning algorithms such as Convolution Neural Network (CNN), Visual Geometry Group (VGG-16) and Artificial Neural Network (ANN) will also be used for the image datasets, and for that datasets the SVM and Naive Bayes algorithms will be used. The hybrid model was created using various ensemble techniques, such as max voting, in which different algorithms were applied to the same data and the result was taken based on the highest same value, averaging, in which after applying the model to the data, the test results were averaged and stored in a new column, and weighted average, which was similar to averaging in which a specific weight is given to each algorithm and does the averaging parsing. And also, most predictions will use the straight forward approach firstly data preparation, which includes data pre-processing in which the data is normalized, resized, and unwanted features are removed that is followed by feature extraction, which consists of useful features and deletes the remaining features and finally, various algorithms are trained on the particular dataset and feed the data to the model, and it is built by running the epoch. The model was fed by testing data and the results were confirmed using multiple performance measures such as Precision, Recall, F1 Score, Accuracy, Right Mean Square Error and many more. The model is subsequently saved into the system via the pickle file and passed to the backend work of the user interface in which the model was run. The user interface was created using current web languages such as react and bootstrap.

2. Literature survey

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019)

The makers of this paper present a novel strategy that combines a linear model and a random forest to diagnose the disease in a specific individual (HRFLM). The radial bias function divides the dataset into training and testing portions of 70% and 30%, respectively. This specific model was compared

to other machine learning models to demonstrate that it was achieving higher accuracy than existing models. Models used for comparison include k-nearest neighbours, decision trees, evolutionary algorithms, and naive bayes. The authors created new hybrid models by combining various machine learning techniques, with the HRFLM having the highest accuracy, and also established a system for computer-aided decision-making, which is utilized to improve accuracy. Additionally, particle swarm optimization was included and is used to improve accuracy. The suggested work is also more accurate than the current models. The dataset was acquired from the UCI repository, and techniques like apriori, predictive, and tertius association rules were utilized in the HRFLM. Additionally, different hybrid algorithms are being used in the prediction by merging the different machine learning methods used based on the results the HRFLM gave more accuracy than the other hybrid algorithms.

[2] Bertsimas, D., Mingardi, L., & Stellato, B. (2021).

In this article, the author provides a framework for using research they conducted on the nodes of the heart disease to determine whether or not a person has the condition. The majority of the information was gathered through ECGs, which have more than 40 thousand and may diagnose diseases within a minute. There are various methods for choosing the features and data, training the model, and applying the model to the data. They used 4 datasets for the model, and from each dataset, they chose a few features, grouped those features, and then extracted features from those grouped datasets. Only 140 of the almost 250 characteristics that were chosen by the ada boost method were ultimately trained using that technique, and Optuna optimization framework was also employed to fine-tune these data. This determines the maximum and minimum number of times that specific algorithm should run on it in order to prevent overfitting the model to the data. The calibration factor, which is nothing more than the predictions being divided into M interval bins and calculating the accuracy of each bin, was used to test the model after it had been trained in order to determine its accuracy and f1 score. Additionally, each dataset's accuracy and f1 score are embellished, and the model's accuracy was 93%, surpassing all previous research in this area.

[3] Shah, D., Patel, S., & Bharti, S. K. (2020).

The authors of this research proposed a brand-new technique called the hybrid method, which may determine whether or not a specific person has lung illness. Additionally, this was accomplished using a variety of machine learning methods, including the random forest algorithm, decision trees, K-nearest neighbours, and naive bayes. The UCI repository was used to obtain the dataset, which has 500 rows of data and also missing values are eliminated from the dataset using data mining techniques. This helps to extract the hidden information from the dataset as well as the relationships that are there in it. After the missing values were removed, training and testing were separated. Then, after these algorithms have been tested on the specific data and trained on it, the results are taken so that the algorithm with the highest precision was coupled with the algorithm with the highest f1score to create the hybrid algorithm. The model was again trained and evaluated, and it provided the best accuracy.

[4] Katarya, R., & Srinivas, P. (2020, July).

This study contrasts the various machine learning models now in use with a model that is being presented that uses deep learning techniques. synthetic neural network in the past, angiography, which may be used to detect numerous disorders at once, was used to forecast heart disease. Additionally, the data collection was compiled from a variety of sources. And in order to develop the specific model, many layers that include various activation functions as well as dropout layers that eliminate extraneous information were incorporated. Additionally, there are a few machine learning methods including support vector machines, random forests, and naive bayes. Additionally, authors discussed the various scientific methodologies that are employed. Therefore, in the future, it is preferable to utilize search algorithms to choose the features, and then machine learning approaches to predict outcomes will provide us better outcomes for heart disease prediction. And with the proposed model, it achieved 93.33% classification accuracy using naive bayes. It outperforms the traditional ANN model by 3.33 percent.

[5] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March).

The paper suggested a system that uses machine learning algorithms which are neural networks because it has been shown to be the most accurate and trustworthy algorithm. The research was initially focused on the primary causes or variables that have a significant impact on heart health. Although certain variables, like age and family background, cannot be changed, others, such as blood pressure and heart rate. The collection of datasets came next. We utilized the Cleveland dataset from the UCI library for this. The collection includes up to 76 factors that describe the total state of the heart's health.an overview of the new sensors on the market that assess several characteristics including the conventional heart. An illness prediction method makes use of 13 key variables offers its consumers a forecast result that reveals a user's current state and leads to CAD. The machine learning algorithms gave great improved result of recent technological breakthroughs; as a result, we adopt Multi Layered Perceptron (MLP) in the suggested system due to its effectiveness and precision.

[6] Yuan, K., Yang, L., Huang, Y., & Li, Z. (2020, November).

In order to increase the precision of machine learning in predicting heart disease, we present a novel technique in this paper called hybrid gradient boosting decision tree with logistic regression (HGBDTLR) that uses ensemble learning. Heart disease data from the Cleveland database is used by the UCI dataset to train HGBDTLR. The UCI data set consists of 303 data points, of which 80% (242 pieces) are separated into training sets and tests, while the remaining 20% (61 pieces) are divided into test sets for classification. The advantage of ensemble learning is embodied by the HGBDTLR algorithm introduced in this research. The BFAHP method was developed by Farnaz Sabahi et al. and demonstrated 87.4% prediction accuracy on the UCI Cleveland dataset. Rough set was proposed to be introduced by Amin et al. On the Cleveland heart disease datasets, a hybrid technique combining linear regression, multiple

adaptive regression splines, and neural networks achieved prediction accuracy of 82.18%, 85.82%, and 91.30%. In order to identify cardiac disease, Wang Jie et al. employed classification models such SVM, Logistic Regression, Naive Bayes, and Naive Bayes and achieved an accuracy of 82%.

[7] Raizada, S., Mala, S., & Shankar, A. (2020, October).

Chikungunya, malaria, and dengue are three infections spread by vectors that have been detected across the Indian subcontinent (Multiclass Classification). The author of this essay uses two different forms of data as the dataset. Demographic data makes up the first of them, while meteorological data makes up the second. Only positive data points are present in the demo graphics, whereas the meteorological data includes information from the meteorological department on temperature, humidity, and rainfall. Supervised machine learning is a type of machine learning. The value given as input is the feature value, X = (x1, x2,..., xn), which combines biotic and abiotic elements. The dataset employed by the author, which was split into 60% training and 40% testing (6:4), included 28 states and 6 union territories of the subcontinent. the methods of machine learning and deep learning. Deep learning and machine learning methods were employed in this. The techniques to anticipate the epidemic include Feed-Forward (FF), Back-Propagation (BP), and Gradient-Descent (GD). the ANN multimodal outbreak prediction (CNN-MDOP) algorithm, which uses data imputation and token conversion. To increase the quality of the data, imputation and tokenization in the data detect and change unclear data. Five layers are employed for prediction in the CNN-Based Multimodal Disease Outbreak Prediction (CNN-MDOP) algorithm. Using CNN, the forecast was 88% accurate.

[8] Li, Z., Xin, J., & Zhou, G. (2022).

The author of this paper has created an integrated spatio-temporal recurrent neural network and nonlinear regression model for the development of vector-borne diseases. A quadratic embedding approach driven by recommendation algorithms encodes the climatic data for the model. A lengthy short-term memory neural network models the impact of nearby areas. The integrated model is evaluated using leishmaniasis data from Sri Lanka from 2013 to 2018 when infection outbreaks occurred and is trained using stochastic gradient descent. By adding spatio-temporal transmission parameters including temperature impacts and local transmission of illnesses from nearby regions, the goal is to generalize and enhance existing geostatistical and ecological models. A recurrent neural network modelling the spread of leishmaniasis between nearby locations using data from the top three affected neighbours. crossbreeding with regression to create an integrated nonlinear space-time model trained by stochastic gradient descent, and using climate data input as an external element because the growth of both sand flies and the parasites in their stomachs are influenced by climatic circumstances. In comparison to ARIMA, which is exclusively based on historical observations in the region of interest, edge features in situations of nearby regions helped the model perform better.

[9] Shimpi, P., Shah, S., Shroff, M., & Godbole, A. (2017, July).

This study suggests a novel approach that uses deep learning to identify three diseases: malaria, dengue fever, and chicken pox. The backpropagation algorithm based on artificial neural networks is used in this particular method. Numerous gradient approaches, including RMSProp, Adagrad, Classical momentum, Adaptive moment Estimation, and Nesterov accelerated gradient, have been applied to boost the suggested method's accuracy. These techniques raise the accuracy up to 99.7 percent. When building the layers of an artificial neural network, the activation function and drop layer were most frequently utilized. Because the frequency of each layer is fixed, we only receive the nodes that are most necessary for the given model. The algorithm has three stages. Phases include Feedforward, Backpropagation of Error and Weight Updating. The suggested strategy goes through these gradient approaches before processing through the model, which is subsequently trained, as well as giving each of the gradient techniques used to train the specific model a separate epoch.

[10] Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December).

A novel method that the researchers had devised can be used to determine whether or not a person has lung disease. Several algorithms, including Support vector machine. Decision trees, Multi-Layer Perceptron's, Naive Bayes, and ensemble approaches like Max voting and cross validation procedures, were employed in this methodology to make this specific prediction. Three specific algorithms were chosen for maximum voting, and these three algorithms anticipate the results depending on the votes. Only 15% of the data showed that the condition could have been predicted at the beginning, in which case the specific person would have received a diagnosis. The UCI repository was used to obtain the dataset. Additionally, it was contrasted with the fuzzy deep learning algorithm in order to compare the outcomes, choosing the best from the two. Moreover, they discovered that the suggested approach may also forecast other diseases like chronic pain, chest discomfort, etc. when forecasting lung disease. For evaluating the output of these models, a variety of performance criteria, including the F1 score, precision, and recall, have been used. The random forest classifier and ensemble approaches are more accurate for these models, with accuracy levels surpassing 80.

[11] Wu, Q., & Zhao, W. (2017, October).

This research introduces a unique technique for identifying small cell lung cancer from computed tomography images, known as the entropy degradation method, which is based on neural networks. The National Cancer Institute provided the dataset. The specific dataset only uses 12 CT pictures for training. The training data is high resolution, and pixels are selected from the images and trained before being tested on 6 more images. The ct images are a sort of 3D image that are broken into sections of 512x512x3 depending on the needs of the model. Convolutional neural networks and other deep learning techniques are used to forecast the occurrence of small cell lung cancer. The prediction is based on both forward and backward propagation, with the

backward propagation using an activation function to boost the model's accuracy. About 512 features from the images are used to train the specific model that can distinguish between the variations in the images. The sole parameter that was used to compare the models was accuracy.

[12] Rahane, W., Dalvi, H., Magar, Y., Kalane, A., & Jondhale, S. (2018, March).

Researchers came up with a novel technique that involves categorizing the disease using blood samples and images. Support Vector Machine, as opposed to other algorithms like naive bayes and k nearest neighbours, provides the highest accuracy for categorization. The ELCAP public lung image database served as the source of the dataset. This has almost 200 training images There were numerous noise images included in the data for these images, and several pre-processing techniques were applied to the images in order to remove that particular noise data. One such technique is grayscale conversion, which is used to extract the image's primary information. The data was then subjected to noise reduction using the median filter to remove undesirable information, followed by binarization, which turns a grayscale image into a binary image, and segmentation, which separates images into classes before being sent to the model-training stage. Before training the model, several features are retrieved from the images based on the region of interest, area, perimeter, and eccentricity. The input was from the output of the segmentation. Then after the model has been trained, the authors have created a graphical user interface (GUI) that primarily uses a SQL database and Java.

[13] Wang, Q., Zhou, Y., Ding, W., Zhang, Z., Muhammad, K., & Cao, Z. (2020).

Lung cancer can be identified early and treated, resulting in a reduction in mortality rates, by applying supervised learning algorithms to train gene expression data. In this study, we present a random forest with self-paced learning bootstrap for enhancing lung cancer categorization and prediction by leveraging gene expressions. The most frequent cancer diagnosed worldwide and the main cause of cancer death is lung cancer. Shortness of breath, wheezing, hoarseness, chest pain, coughing, or spitting up blood are all indications of lung cancer. Early cancer detection can be achieved using machine learning techniques. As a potential tool for cancer categorization and diagnosis, machine learning has developed. We present a unique random forest with self-paced learning (RFSPL) in this study, enabling us to efficiently extract high-quality data from the lung microarray data. Using DNA microarray technology, the RFSPL framework is used to train a model for cancer categorization and diagnosis. In terms of accuracy, F1-score, and area under the curve, our suggested technique outperforms currently available classifiers (AUC). Data on cancer prognosis can be classified using random forest, which can solve this challenge. In comparison to the other approaches, RFSPL can obtain higher accuracy, AUC, and F1-score values for each class. Both highland low-quality samples can teach the SPL new things. Other disease diagnoses can also be obtained using different bagging techniques. The semi-supervised classification is used since it is simple to gather numerous unlabelled samples.

[14] Kumar, M. V. (2020, July).

This document summarizes studies on lung disease identification from numerous papers in the field of machine learning. The study uses a variety of models, most of which are drawn from deep learning and machine learning techniques. Additionally, the author proposed his work on the detection of lung diseases using a convolution neural network. However, this method had some limitations, such as its inability to detect diseases in images smaller than 5 mm. As a result, numerous upgrades were made to the convolutional neural network, which consists of many kernels, as well as multidimensional models, in order to detect these diseases. Only 350 high-resolution images from CT scans made up the suggested work dataset, which was divided into three categories: training, testing, and validation. Additionally, suggested work using the conventional approach and the convolution method was completed. The classic approach uses image processing to reduce the size of the image node by node before training the model in a step-by-step manner. whereas the convolution method trains the model while also resizing it at the layers level. Convolution has higher accuracy than the other two approaches out of these two.

[15] Mohalder, R. D., Sarkar, J. P., Hossain, K. A., Paul, L., & Raihan, M. (2022, March).

In this study, a brand-new work that only uses machine learning techniques for classification was introduced in the publication. The algorithms employed in this work are Light GBM, XGBoost, Naive Bayes, K-nearest Neighbour's, Support vector machine, and Random Forest. compared these algorithms against other algorithms, as well. Additionally, this work evaluates the other paper works and concludes that the suggested work is more accurate than the others. The dataset, LC25000, included lung pictures that might be used to categorize the five disorders. Resizing, separating, training the model, and finally validating it were the first steps. Additionally, a thorough explanation of each algorithm's operation was provided. The results are then obtained after visualizing each model, and the training period and accuracy of each model are also tabulated. Different performance criteria, such as precision, recall, F1 score, and support, are employed to assess the model. According to these tables, Light GBM has the highest accuracy.

[16] Aykanat, M., KILIÇ, Ö., Bahar, K. U. R. T., & SARYAL, S. B. (2020).

The study on lung disease that can determine whether a specific person has the disease or not is presented in this paper and consists of both hardware and software implementation. In this specific disease, Support vector machine (SVM), Naive Bayes, and K-Nearest Neighbours were the techniques employed. This model was developed using a variety of datasets, including 12 diseases with text data and 12 diseases with audio MFCC features, as well as sick or healthy with text data, sick or healthy with audio MFCC features, and sick or healthy with text data and audio MFCC features. The research also reveals the existence of a developing model, known as the stethoscope, which is useful for sensing the object on which it was trained using algorithms. Only accuracy was utilized as the metric for evaluating the algorithms. The highest accuracy is only achieved by the support vector machine.

[17] Ahsan, M. M., Uddin, M. R., Farjana, M., Sakib, A. N., Momin, K. A., & Luna, S. A. (2022).

The study explains how to gather information on monkeypox from various sources and train a specific model to determine whether or not an individual has the disease. The Visual Geometry Group (VGG-16) is utilized in this study for training, and Local Interpretable Model-agnostic Explanations (LIME) is employed to examine the model that underlies the final predictions. Data augmentation, which is used to expand the data by zooming the data, rotating the image, and other methods, was done since the data was relatively low. After that, it was supplied for training, and LIME entered the picture to collect data from the super pixels. Different layers are employed in the model to shrink the data and remove any unnecessary information from it. The loss of the data was managed using the Adam optimizer. The model is tested using a variety of performance metrics, including precision, recall, accuracy, sensitivity, specificity, F1 score, and area under the curve. Due to the fact that there was only one method, it was performed on many epochs before being compared and tabulated.

[18] Ali, S. N., Ahmed, M., Paul, J., Jahan, T., Sani, S. M., Noor, N., & Hasan, T. (2022).

Images with dimensions (224, 224, 3) were fed into pre-trained models. The bottom eight layers were unfrozen to ensure homogeneity and better generalization. An adaptive learning rate optimizer with an initial learning rate of 10 −5 was employed for training. ResNet50 had the best accuracy (82.96±4.57%) and VGG16 had the most competitive performance (81.48±6.87%) of the 3 models. System can be further improved by using a multi-source dermato- scopic image dataset for pre-training the models. The dataset is created primarily by web-scraping, which lacks meta-data that is vital for diagnosis. A more concerted effort and international collaboration is needed to collect a larger dataset. Deep learning architectures (VGG16, ResNet50, InceptionV3) leveraging the transfer learning approach. Despite being a small dataset, the promising results obtained after 3-fold cross-validation reveal the potential to use AI-assisted early diagnosis of this disease.

[19] Iradukunda, O., Che, H., Uwineza, J., Bayingana, J. Y., Bin-Imam, M. S., &NiYonzima, I. (2019, December).

In this paper, a novel approach called extreme machine learning was employed to identify the presence of malaria based on images. Additionally, the elm was contrasted with other machine learning algorithms, including support vector machine, k nearest neighbours, random forest, residual networks, convolution neural networks, and dense net. It follows a specific workflow throughout the project. The Lister Hill National Centre for Biomedical Communications dataset was used in this study. The first step is pre-processing, which consists of three steps: grayscale image conversion to improve image quality, normalizing to ensure that image sizes are consistent across the dataset, and finally, image augmentation to improve detection rate. Then the following phase in the procedure is feature extraction, which uses the Hu-moment approach and harlick feature extraction to crop the image to the area where the specific spot was present. Followed by our model, also known as feedforward neural networks, that is Elm. 3 layers make up this elm (input layer, hidden layer, output layer). The white box, often known as the elm, is made up of several concealed layers. In these layers, activation functions are used to increase stability. The dense net was used as the primary comparison, and the suggested model provided greater accuracy in less processing time.

[20] Militante, S. V. (2019, December).

In this paper, the author evaluates and contrasts several CNN models. After examining the results from the various models, ResNet, Google Net, and VGGNet provided excellent accuracy ranging from 90% to 96%. The application of a computer-aided diagnostic (CADx) tool to determine if a certain picture was infected or not required human work, but the work also yielded superior results. The author also addressed the major deep learning models and algorithms, primarily VGG-16, ResNet, and Google Net. further demonstrated how each model's internal workings operate. The workflow model has four steps. Data gathering and data analysis come first, followed by image pre-processing, which includes scaling, cropping, and histogram. The following step is training, which actually involves feature extraction, fine-tuning, and finally testing the model. Compared to these 3 models VGG-16 model has the higher accuracy.

[21] Oyewola, D. O., Dada, E. G., Misra, S., & Damaševičius, R. (2022).

The author introduces a unique deep learning model called data augmentation convolutional neural network that is taught via reinforcement learning in this paper. In addition, the performance is compared to that of the convolution neural network. The dataset was obtained from the Kaggle database, which contains 27588 images. The convolution layer functions as a feature extractor, extracting just the pixels of the picture that contain the real disease and ignoring the rest. The directed acyclic graph convolution neural network convolution neural network convolution neural network was used to create the graph-like structure in which the Data Augmentation of Convolutional Neural Network layers will be present at the nodes and these layers will be able to train and test the data. Data augmentation was performed at each node to improve model accuracy, and this data augmentation was performed by rotating the data at that specific location by +20 deg or -20 deg. To assess the data, many performance indicators were employed, including Mean Absolute Error, Root Mean Square Error, Mean Absolute Scaled Error, Accuracy, Specificity, Sensitivity, Kappa, and detection rate. After analysing the results, it was determined that data augmentation convolution neural networks exceeded convolutional neural networks with greater accuracy.

[22] Qin, B., Wu, Y., Wang, Z., & Zheng, H. (2019, October).

In this study, an evolutionary convolutional deep network (ECDN), which can build its own deep neural networks and optimize their topology throughout the evolution process, was developed as a data-driven technique for detecting malaria. By developing the ECDN model, it can create its own architecture,

which cuts down on time and errors. Such that it may also be directly connected to any keras-compatible deep neural network platform. The collection, which consists of 27558 images, is taken from the National Institutes of Health (NIH). The initial step was pre-processing, which includes sample purification, image rescaling, and data enhancement, in which the incorrectly marked data is removed, the image is resized to a specific scale, and the image is rotated, mirrored, and cropped to increase its diversity. Authors had used histogram equalization approach to reduce noise in the photos and YUV-Space to boost the overall contrast and brightness equivalence of the image. The following is our model ECDN, where we initially established the population size before defining our fitness function to repeatedly iterate till our condition was met. The model has eight layers, including kernel and max-pooling. The 6:4 ratio, which had provided 99.98 percent in detection, was chosen as the best split when the accuracy was compared with the various splits. Additionally, the model was assessed against others like the VGG-16 and CNN and our model produced unexpected results.

[23] Shah, D., Kawale, K., Shah, M., Randive, S., & Mapari, R. (2020, May).

This paper proposes a new methodology for identifying malaria. That is both faster and less expensive. The Convolution neural network model was used, and it only has three fully connected layers. And this CNN performed admirably with less resources while providing the highest accuracy. This study includes pre-processing, training the model, testing the model, and evaluating the performance metrics. The pre-processing includes image recognition and image classification, which aid in resizing the image and reducing noise and image contrast. Following that, the images are labelled to determine if they are uninfected or parasitized. After that, the appropriate classifier selects the features, and the features are derived from the images, such as histograms. Finally, the image data generator was employed to provide additional variability in images throughout the training step. The CNN model was then executed, which has three layers, the first two of which use the RELU as an activation function, 32 and 64 filters in the first two levels, and 128 filters in the last layer, which employs the sigmoid function. And the model was tested using previously unknown photos, and predictions were made using these images. For all six epochs, the model provided a consistent accuracy rate of 95%.

[24] Olugboja, A., & Wang, Z. (2017, July).

In this article, a quick and accurate technique was created using images of a blood smear. Additionally, the authors used the watershed segmentation technique to locate the plasmodium's infected regions. Moreover, six machine learning algorithms—Linear SVM, Quadratic SVM, Fine Gaussian SVM, Cosine KNN, Boosted Tree, and Subspace KNN—are implemented for classification. They used a simple method for identifying infected cells in images, which included pre-processing the images to make them grayscale for better visualization, segmenting the images into regions of interest and non-regions of interest, and using watershed segmentation to identify infected cells. The built-in functions of MATLAB were used to compare the results. Based on the criteria of accuracy, true positive rate, and false negative rate, the subspace KNN and fine gaussian SVM had performed better when classifying photos with Plasmodium and Babesiosis infection.

[25] Yadav, S. S., Kadam, V. J., Jadhav, S. M., Jagtap, S., & Pathak, P. R. (2021, March).

An experimental analysis of several machine learning models to detect malaria disease has been done in this paper. In this study, there were two datasets. From those two datasets, three more datasets were built by removing some factors, combining the two datasets, and extracting common attributes. Several algorithms, including Random Forest, Support Vector Machines with Gaussian Kernels, Artificial Neural Networks, Naive Bayes, and Logistic Regression have been applied. Follows a fundamental approach to data preparation, which includes cleaning and normalizing the data as well as identifying and replacing any missing values. After these, all algorithms with the identical parameters were applied to each dataset. To analyse the performance of these algorithms, some performance metrics including as precision, recall, f-measure, specificity, and roc are used. From these results, the random forest, linear regression, support vector machine with gaussian kernel, and artificial neural network produced better results. The artificial neural network gave unexpected results.

3. Conclusion

To bring our discussion to a close, we examined other articles in which researchers had finished their work on various models, as well as with related technologies and their accuracy., we reviewed various diseases like Heart, Lung, Monkeypox, Vector Borne and Malaria and various machine learning algorithms are been used by the researchers such as support vector machine, naive bayes, K-nearest neighbor, random forest, decision tree, and some deep learning algorithms such as convolution neural network, artificial neural network. The most current publications were chosen for the study, which will be published in Science Direct, IEEE Xplore, Elsevier, and ResearchGate indexed journals and conferences until June 2020.Most of the studies undertaken for illness prediction used datasets from Cleveland, UCI repository, Lister Hill National Center for Biomedical Communications, and very few used their own datasets. Analyzing the existing methodologies used for illness prediction reveals that, while certain methods have made a substantial increase in results with high sensitivity or less FP, there are still numerous issues to be solved. As a result, an optimum approach for predicting these diseases as early as feasible is still required, and this review study article will be useful for researchers and professionals. While some strategies have significantly improved findings with high sensitivity or less FP, an analysis of the current methodologies used for sickness prediction demonstrates that there are still many problems to be resolved. This review study article will be helpful for researchers and professionals in finding the best strategy for identifying these diseases as early as possible.

4. References

[1]. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[2]. Bertsimas, D., Mingardi, L., & Stellato, B. (2021). Machine learning for real-time heart disease prediction. IEEE Journal of Biomedical and Health Informatics, 25(9), 3627-3637.

[3]. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science,1(6),1-6.

[4]. Katarya, R., & Srinivas, P. (2020, July). Predicting heart disease at early stages using machine learning: a survey. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 302-305). IEEE.

[5]. Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1275-1278). IEEE.

[6]. Yuan, K., Yang, L., Huang, Y., & Li, Z. (2020, November). Heart disease prediction algorithm based on ensemble learning. In 2020 7th International Conference on Dependable Systems and Their Applications (DSA) (pp. 293-298). IEEE.

[7]. Raizada, S., Mala, S., & Shankar, A. (2020, October). Vector borne disease outbreak prediction by machine learning. In 2020 International conference on smart technologies in computing, electrical and electronics (ICSTCEE) (pp. 213-218). IEEE.

[8]. Li, Z., Xin, J., & Zhou, G. (2022). An integrated recurrent neural network and regression model with spatial and climatic couplings for vector-borne disease dynamics. arXiv preprint arXiv:2201.09394.

[9]. Shimpi, P., Shah, S., Shroff, M., & Godbole, A. (2017, July). An artificial neural network approach for classification of vector-borne diseases. In 2017 International Conference on Computing Methodologies and Communication (ICCMC) (pp. 412-415). IEEE.

[10]. Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December). An evaluation of machine learning classifiers and ensembles for early-stage prediction of lung cancer. In 2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST) (pp. 1-4). IEEE.

[11]. Wu, Q., & Zhao, W. (2017, October). Small-cell lung cancer detection using a supervised machine learning algorithm. In 2017 international symposium on computer science and intelligent controls (ISCSIC) (pp. 88-91). IEEE.

[12]. Rahane, W., Dalvi, H., Magar, Y., Kalane, A., & Jondhale, S. (2018, March). Lung cancer detection using image processing and machine learning healthcare. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-5). IEEE.

[13]. Wang, Q., Zhou, Y., Ding, W., Zhang, Z., Muhammad, K., & Cao, Z. (2020). Random forest with self-paced bootstrap learning in lung cancer prognosis. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1s), 1-12.

[14]. Kumar, M. V. (2020, July). Detection of lung nodules using convolution neural network: a review. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 590-594). IEEE.

[15]. Mohalder, R. D., Sarkar, J. P., Hossain, K. A., Paul, L., & Raihan, M. (2022, March). Efficient Machine Learning Techniques to Predict Lung Cancer. In Proceedings of the 2nd International Conference on Computing Advancements (pp. 233-239).

[16]. Aykanat, M., KILIÇ, Ö., Bahar, K. U. R. T., & SARYAL, S. B. (2020). Lung disease classification using machine learning algorithms. International Journal of Applied Mathematics Electronics and Computers, 8(4), 125-132.

[17]. Ahsan, M. M., Uddin, M. R., Farjana, M., Sakib, A. N., Momin, K. A., & Luna, S. A. (2022). Image Data collection and implementation of deep learning-based model in detecting Monkeypox disease using modified VGG16. arXiv preprint arXiv:2206.01862.

[18]. Ali, S. N., Ahmed, M., Paul, J., Jahan, T., Sani, S. M., Noor, N., & Hasan, T. (2022). Monkeypox skin lesion detection using deep learning models: A feasibility study. arXiv preprint arXiv:2207.03342.

[19]. Iradukunda, O., Che, H., Uwineza, J., Bayingana, J. Y., Bin-Imam, M. S., & Niyonzima, I. (2019, December). Malaria disease prediction based on machine learning. In 2019 IEEE international conference on signal, information and data processing (ICSIDP) (pp. 1-7). IEEE.

[20]. Militante, S. V. (2019, December). Malaria disease recognition through adaptive deep learning models of convolutional neural network. In 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS) (pp. 1-6). IEEE

[21]. Oyewola, D. O., Dada, E. G., Misra, S., & Damaševičius, R. (2022). A novel data augmentation convolutional neural network for detecting malaria parasite in blood smear images. Applied Artificial Intelligence, 1-22.

[22]. Qin, B., Wu, Y., Wang, Z., & Zheng, H. (2019, October). Malaria cell detection using evolutionary convolutional deep networks. In 2019 Computing, Communications and IoT Applications (ComComAp) (pp. 333-336). IEEE.

[23]. Shah, D., Kawale, K., Shah, M., Randive, S., & Mapari, R. (2020, May). Malaria parasite detection using deep learning:(Beneficial to humankind). In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 984-988). IEEE.

[24]. Olugboja, A., & Wang, Z. (2017, July). Malaria parasite detection using different machine learning classifier. In 2017 International Conference on Machine Learning and Cybernetics (ICMLC) (Vol. 1, pp. 246-250). IEEE.

[25]. Yadav, S.S., Kadam, V.J. Jadhav, S., Jagtap, S., & Pathak,P.R(2021,March).ML based malaria prediction using clinical findings. In 2021 International Conference on Emerging Smart Computing and Information (ESCI), IEEE.