Noname manuscript No.
(will be inserted by the editor)

Anomaly Detection In Biomedical Data and Image Using Various Shallow and Deep Learning Algorithms

Md. Nurul Absur

Received: date / Accepted: date

Abstract Data science has become a hot buzzword in the current world. Analysis of big data with the help of shallow and deep learning methodologies has been increased in computer science, education, business technologies, medical science, manufacturing engineering, CRM, bioinformatics, statistics, etc. It has excellent potential to develop the healthcare system. Various multi-dimensional data sets, data visualization techniques, machine learning, and different data mining approaches are used extensively by scientists to predict patients' conditions and reduce the cost of ensuring that the patients get appropriate treatment at the right time and in the right place. Implementing different shallow and deep learning approaches to the medical field reduces medical errors and ensures patient's reliability and safety. Besides, if these methodologies can be helpful ineffective treatment, which may reduce the cost of unwanted medical tests, medication errors can also be kept at a tolerable level, and medical errors can be reduced. In this paper, the effective use of shallow and deep learning and their performances in various circumstances is portrayed with relevant results. Practical usage of machine learning algorithms will pave the way to produce better work in medical science and sublime the treatment procedure in terms of effectiveness.

Keywords Deep learning · Computer vision · Medical data · Machine learning · Cognitive intelligence · Healthcare

1 INTRODUCTION

In the modern era, intelligent assistance decision-making and prediction systems help doctors and other medical professionals diagnose diseases such as heart disease, diabetes, hepatitis, chronic kidney diseases, etc. The prime

---

Md. Nurul Absur
Department of Statistics, Jahangirnagar University, Dhaka, 1342, Bangladesh
E-mail: nurulabsur11@gmail.com

factors, in this case, are stability, performance accuracy, prediction accuracy. If these factors are not up to the mark, then it becomes challenging relying upon the intelligent system. It may lead to a low-quality prediction system which may ruin many patients' lives. The increasing number of data in the biomedical domain nowadays leads to the need for data analysis. It has a potential impact on predicting diseases intelligently. Computational intelligence with shallow and deep learning plays a vital role in locating patterns, relationships, and knowledge among medical data sets.

Heart disease, Diabetes, Hepatitis, etc., are the most common diseases around the world. Heart disease is the leading cause of death all over the world for the last few decades. Many people whose age is above 40 are affected by Diabetes. The European Public Health Alliance reported that the reasons behind the 41% deaths are heart strokes, heart attacks, and many other cardiovascular chronic diseases. "The Economical and Social Commission of Asia and the Pacific claimed that most lives in one-fifth of the Asian countries had been lost to heart, cardiovascular diseases, Diabetes, etc. The Australian Bureau of Statistics claimed that almost 33.7% of all deaths are caused by heart and other circulatory diseases, and it's the leading cause of death. In the case of South Africa, the statistics of that country reflect that heart diseases are the third leading cause of death [19] .

Studying the background of the leading causes of deaths worldwide and the availability of databases and data sets in the medical field have led the researchers to apply different classification techniques of shallow learning -, deep learning methods in the medical areas to assist the medical professionals. Using these techniques has become beneficial in this field to increase the efficiency rate. Many researchers have found that hospitals worldwide don't maintain the same quality of service, but the types are all the same. Hence, the reports that are being generated are almost the same in most hospitals. They have proposed several suggestions to identify effective treatments so that the performance of the practitioners may be improved. Researchers also suggest that applying machine learning and deep learning techniques in the medical field can provide more reliable accuracy [10].

Our contributions in this paper are-

- Our approach will help any biomedical professionals to describe and analyze the pattern of disease.
- Our proposed image classification model uses a deep learning approach to detect seven types of disease from the frontal x-ray images.
- The proposed approach will help determine what kind of learning reflects better results in different disease circumstances.

II. RELATED WORKS

Over the years, computer vision plays a vital way in predicting and determine the different states of disease in both human and animal bodies. The precise use of algorithms in relevant data sets can produce tremendous successive accurate results. As data science has developed rapidly in recent years, Deep Learning has been introduced to determine the medical state of instances very effectively from the image. This particular approach is a relatively new but convenient approach to data science. Throughout the years, a lot of researchers and authors discuss data science in medical science.

Igor Kononenko talked about subareas in medical image classifiers and their history [17] . He described the history of medical data processing from an image from the Naïve Bayes, Decision Tree, and Neural Network classification model approach. Though his works were not recognized fully by medical science, they helped many future researchers in this field. Wernick et al.proposed a model with Support Vector Model assistance which was quite relevant to medical image processing in medical science [30] . He also discussed the differences between SVM and RVM methods. Both are kernel methods, so their differences need to be understood for future work. Clinical Variables is another parameter system to determine and predict disease from image samples. Wang and his team also proposed a similar approach [29] . He divided the tree tissue data set into subsamples to acquire better instances. The regression of RVM was the backbone of his postulates. Gray matter, blue matter, and CSF were the most critical parameters that were used in the experiment. In recent years, Deep Learning is the most popular approach to solving image conversion and predicting any disease. Litijens et al. wrote an article on how deep learning works differently in various approaches [20] . They also surveyed how the image analysis process should be adopted for better results with Deep Learning.

In this modern era, medical science wants to be précised about patient health and prospect. Many researchers suggested that ML and Deep Learning need to be merged with medical science to better the patient and associated individuals. From this observation, Greenspan and his team proposed a paper discussing the overview and future works in this field. They mainly focused on Convolutional Neural Network (CNN) and its effect on the medical data set [13] . They also conducted a test set over various instances to determine if the algorithm works the same for all data types. They found out that the result is optimum in maximum cases, and CNN is the perfect choice for the classification. Gulshan V. et al. took 5560 image samples for their experiment where all the images are from the retinal fundus. The output of this experiment is very much appreciable for its better accuracy, which is near to 97 percent. They eliminated all possible noisy data and counterfeit samples to make it more efficient. Another vital contribution in this field is from Sirinukunwattana and his team about colon cancer [14]. This is locality-sensitive data that was trained and tested by Deep Learning.

Heart Disease is one of the leading causes of morbid- ity and mortality in the world right now [25]. “Survival” is usually defined as the state of a patient remaining alive for a specified period after the detection and diagnosis of the disease. The term was also used in various historical medical testaments. Generally, these types of conventional statistical techniques such as the Kaplan-Meier test and Cox-Propositional hazard models were used to representing the state of the patient **(year?)**. SVM algorithm is a very popular approach that was adopted by various analyzers [9] [28]. The Support Vector Model is very much popular in different cases of the experiments.

SVM classifier model has its individuality which helps to validate training and testing datasets for better prediction. ANN architecture is another popular approach that was tested previously to detect accurate heart disease state [12]. Fuzzy logic is also very notable in this sector. The accuracy of Fuzzy logic was quite impressive. Implemented by N.Bhatla and k. Jyoti., Hornik, et al. empirically showed that the MLP is capable of learning arbitrarily complex nonlinear functions to arbitrary accuracy levels [31] [3].Fuzzy logic has discrete mathematical abilities which help every researcher to decide on the test set. Another contribution based on chronic kidney disease has achieved cult status in this sector. Chronic kidney disease is a significant and lethal disease in recent times. Moreover, the diversity of kidney disease is alarming for every single individual [16]. Chronic kidney disease is a vastly occurring health crisis. For example, in the year 2005, there were approximately 58 million deaths worldwide, with 35 million attributed to chronic disease, according to WHO (World Health Organization) [22]. Many researchers adopted WHO's results in recent years to contribute to the human sector of improvement.

Dhayanand et.al have used SVM and ANN techniques to determine the accuracy of detection regarding chronic kidney disease [18]. ANN is also known as Artificial Neural Network, which is a new concept to work with ease. Predicting survival time was another approach that was quite prospective at the time implemented by Dixon. 11.Db/Db mouse have imple- mented by Kumar's. [24].This method is still popular because of the excellent implementation rate.M.M.Sharanya has worked with three different algorithms to determine which approach is best to predict disease [23]. Neural Network is another approach used by various scientists [27].

III. DATA DESCRIPTION

A. Data for Shallow Learning

This paper reflects that various algorithms work extensively for predicting the biomedical data of Heart Disease, Heart Statlog, Hepatitis, and Diabetes. By using these algorithms, a machine can predict the most suitable algorithms to determine the disease properly by analyzing the accuracy level [26].

The first step is to prepare a medical data set that will use to develop the predicting model. For this application, I gathered workable instances of heart disease, heart Statlog, hepatitis, and diabetes data set from the UCI repository. However, some data was missing and not organized. So, the missing values were filled up by taking the mean of the other values and putting that value on to the missing spot.

In the database of heart disease, there were 75 attributes. Of those 75 attributes, 13 traits were kept, and another 62 features were discarded. There was a total of 303 values. Among them, two values were missing. So, those two values were ignored, and 301 deals were kept for work purposes.

In the Heart data set, there are 13 attributes—a total of 270 values and no missing values.

In the Hepatitis data set, there are 19 attributes. Total 155 values, and there were a few missing values. The process filled the missing data using the meaning of the other values.

In the Diabetes data set, there are 20 attributes. Total 768 values, and there were a few missing values. The process filled the missing values using the means of other values.

1.1 Data for Deep Learning

The first step was to prepare the data set that we will be using to predict different biomedical diseases. We gathered X-ray images of different patients having various conditions. The data that was collected had different dimensions and different aspect ratios. Initially, we cropped the pictures and had them in an ordinary resolution. We also used scalability properties in our images to scale the images for different resolutions. This process improved the model's accuracy when testing with the noisy image.

In deep learning, we need a massive amount of data to train a particular classifier. However, medical image data is confidential and is not readily available. Most of the patients don't want to give their medical imaging reports to the public. Hence it is complicated to create an error-free classification model. I took a data set of frontal x-ray images from "NIH Clinical Center." [32].

"NIH Clinical Center" has the largest database of frontal x-ray images. All the images were in JPG format. Among all the other repositories, the photos were not in JPG format. The data set we took was already in a standard aspect ratio. So, it was easier to pre-process the images.

The image dataset used for this contribution consists of 4999 images of frontal x-ray. Among the 4999 images, it contains 14 Thorax diseases such as Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia. I classified the images from these 4999 images and took 1300 images that was successfully organized. We organized the image data set for 7 Thorax disease that we could identify as Atelectasis, Cardiomegaly, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax—sequentially checked for different aspect ratios and noise in the image. After that, we cropped some images so that all the training images had a standard aspect ratio.

At first, I prepared the Atelectasis thorax disease dataset. Firstly, I trained around 150 Atelectasis disease data sets. Atelectasis is a collapse or incomplete expansion of the lung. For this reason, the lung collapses on both sides or any one side. By using the frontal x-ray image, this characteristic can be easily identified for classification.
In the respective figure, the right lung didn't expand as the left lung. These types of images have been trained as Atelectasis.

Secondly, we trained the Cardiomegaly data set. Here we have trained about 250 images. In Cardiomegaly disease, the heart is enlarged, caused by either coronary artery disease or high blood pressure.

Thirdly, I trained Mass data sets. Mass is a spot in the chest x-ray which indicates there are some problems in the lung. Usually, depending on the shape and size of the mass, lung cancer is detected.

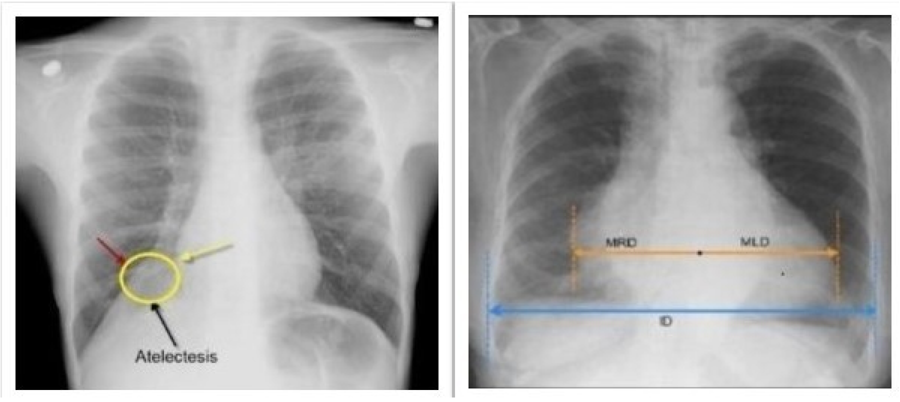

Fig. 1 1. (a) Atelectasis (b) Cardiomegaly

The Nodule is also similar to the Mass, but the size of the Nodule is small. The Nodule is seen more frequently in a chest x-ray. In every 500-chest x-ray image, we can find one Nodule. Based on the characteristics of the images, the Nodule data set was trained and added to the training set.

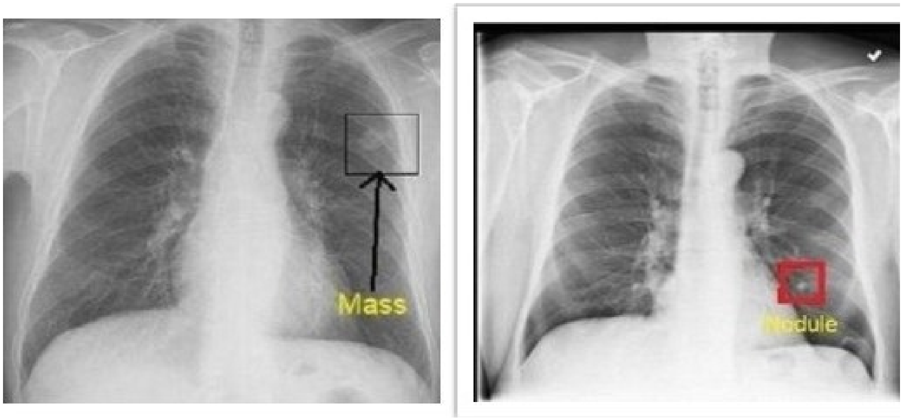

Fig. 2 2. (a) Mass (b) Nodule

I trained the Pneumonia data set for the preparation of a deep learning approach. Pneumonia is a condition where the lung is inflamed due to the air sacs fill with pus. So, the lung is enlarged. In Pneumonia, the air sacs are inflamed and filled with pus. I detected these images and classified them into the Pneumonia data set.

Finally, I trained the Pneumothorax data set. It is a condition where the air in the cavity between the lungs and the chest wall causes lung collapse.

I applied Convolutional Neural Network (CNN) over those images. With the help of Tensor Flow and implementing the inception model, a method

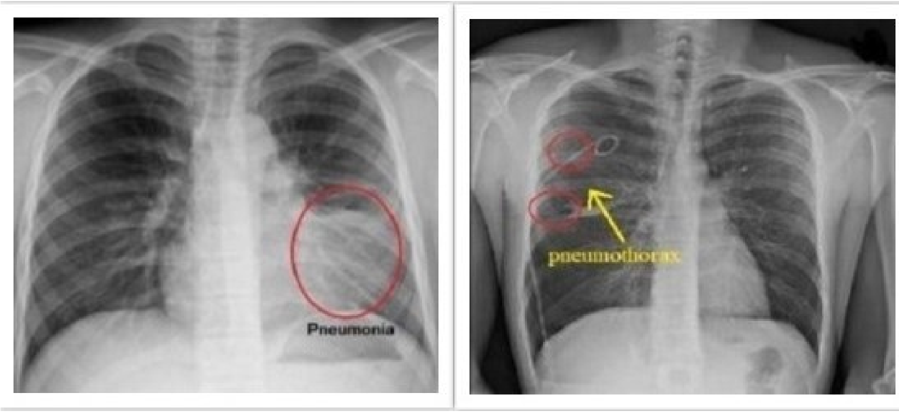

Fig. 3 3. (a)Pneumonia (b) Pneumothorax

was developed, which is an image classification model that classifies the input images [2].

IV. EXPERIMENT SETUP

A. Shallow Learning

The first step is to prepare the medical data sets that will be used to develop predicting models. I gathered data on Heart disease, Heart Statlog, Hepatitis, and Diabetes from the UCI repository. However, some data was missing and not organized. So, the missing values were filled up by taking the mean of the other values and putting that value on to the missing spot.

The heart disease dataset consists of 75 attributes. Of those 75 attributes, 13 features were kept, and the other 62 attributes were discarded. There was a total of 303 values. Among them, two values were missing. So, those two values were ignored, and 301 deals were kept for work purposes [5].

In Heart Statlog, there are 13 attributes. A total of 270 values and no missing values [7].

In the Hepatitis data set, there were 19 attributes. A total of 155 values, and there were a few missing values. The process filled the mislaid data using the mean value of the other values [6].

In the Diabetes data set, there were 19 attributes. A total of 768 values, and there were a few missing values. The process filled the irretrievable values using the mean value of other values [8].

The proposed methods are currently using WEKA version

3.8 software to use machine learning algorithms to detect anomalies over these [4] data sets. Naïve Bayes (NB), Bagging, Random Forest, Logistic Regression, and Random Tree are five classification algorithms used in this study. Four phases are involved (data set planning, learning, validation, and testing). Second, by loading data sets from the network that need to identify data instances. The second stage (training) continues once the data set is ready. Choose the algorithms used in this experiment (as mentioned earlier). The algorithm parameters used are set to default as in WEKA. Second, there is a high degree of reliability in ten-fold cross-validation, even though there are sparse details where knowledge is not wasted—eventually, the evaluation stage collects the output measurements (precision and time taken).

1.2 Deep Learning

The Data set had about 13,000 images. There were about 2% of the images have some noise. We found out and discarded those images. After that, we categorized the X-ray images according to some characteristics.

The second step was to build the inception model. The inception model is the Convolutional Neural Network (CNN). The model consists of three parts:

- Convolution
- Polling
- Flattening

Convolution extracts feature from an input image. Convolution preserves the spatial relationship between pixels. In terms of unit, each image can be considered as a set matrix with pixel values. Pixels are represented as a matrix if they consist of 0 and 1- pixel values. For this experiment, the convolution of a 5×5 image using a 3×3 matrix was computed and achieved the convoluted value.

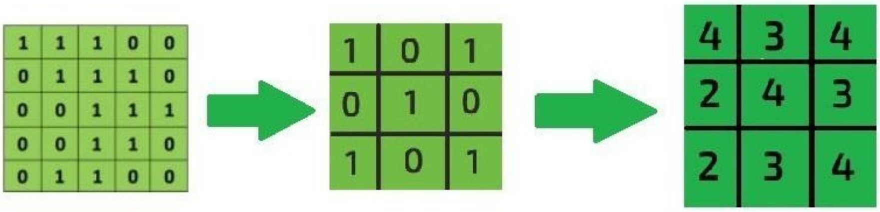

Fig. 4 4. Convolution Process

The next step is polling. Polling is also known as down-sampling. It reduces the dimensionality of each feature map but extracts the essential element. Thus, it gets the required information with less memory usage, and hence it saves the bandwidth. In max polling, a 2×2 window area was defined and took the highest value from that 2×2 window.

The final phase is flattening. This process represents the matrix into a linear array.

As the CNN layer below, the inception model is a combination of 1×1 convolution layer, 3×3 Convolution layer, 5×5 Convolution layer with their output filter bank connected into a single vector that is used as an input for the next stage. The outputs are then fed into the fully connected network. The network is then combined with a convolution neural network.

The inception model can be created by the addition of a 1×1 Convolution layer before applying to another layer which can reduce the dimensionality.

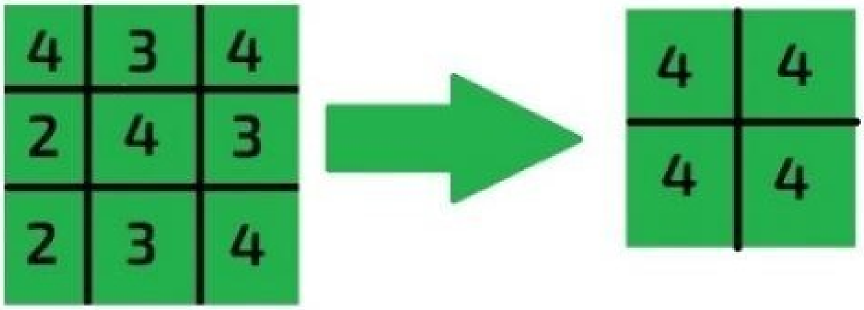

Fig. 5 5. Polling Process

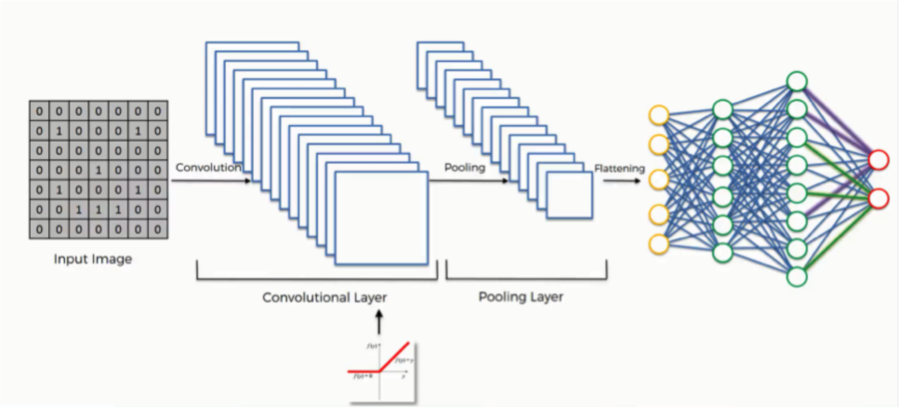

Fig. 6 6. Convolution Neural Network [26].

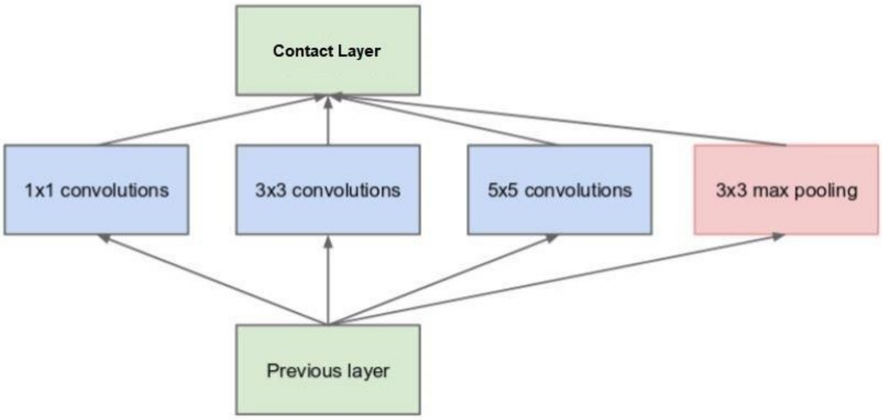

Fig. 7 7. Inception Model

V. RESULT & PERFORMANCE ANALYSIS

A. Shallow Learning

In this section, we will discuss the text data sets which have been used for shallow learning. In our machine learning, we have taken care of four different diseases, and their accuracy regarding various single classifier approaches.

The accuracy of correctly classified elements from this data set represents the success rate of the classifier approach. In terms of detecting disease, a more accurate result is highly appreciated. Usually, accuracy is represented in percentage. The benefit of percentage representation is that it helps to convert any bigger units into smaller ones.

1) Heart Disease: Another widely affected disease is heart disease. Although medical science has reached a better position for heart treatment than other diseases, heart disease still lacks computational access makes it difficult to detect heart disease effectively. This experiment on sample heart disease data set by using single and joint ML classifiers will lead to day-to-day life. For our data set of heart disease, Logistic Regression has better results among the single ML classifiers with 84.48%.

This data set is enriched with 303 instances where thirteen attributes are representing their corresponding values. Five different classes are used to determine the desired output. They are “>50”, ”<50_1”, ”50_2”, ”50_3”. and ”50_4”.

Table 1 Performance Comparison

Classifier Name	Correct Case	Incorrect Case	Accuracy
Bagging	246	157	81.18%
Logistic Regression	256	47	84.48%
Naive Bayes	253	50	83.49%
Random Forest	252	51	83.14%
Random Tree	225	78	74.28%

2) Heart Statlog Disease: This data set is quite similar to the Heart disease data set that we previously used, but it has different forms and different patient instances. These instances are individual and quite compatible enough to work with ML classifiers. This two-class represents two outputs for the ML classifier results. Class-1 describes how many samples are not affected by the disease. Class-2 represents the number of instances that are affected by this disease.

This data set consists of 270 instances with 13 different attributes. The best classification model for this data set is Linear Regression which has 83.70 \% accuracy.

3) Diabetics Disease: The Diabetics disease data set consists of 768 instances divided into two different classes named “tested_ negative” and “tested_ positive,” respectively. There are nine other attributes which are also known as the parameter of the instances.

Table 2 Performance Comparison

Classifier Name	Correct Case	Incorrect Case	Accuracy
Bagging	214	156	79.25%
Logistic Regression	226	44	83.70\%
Naive Bayes	226	44	83.70%
Random Forest	220	50	81.48%
Random Tree	206	64	76.29%

Diabetes is one of the most common diseases in the current time. The proper detection and correct identification can add a new dimension to the medical field. We used five different single classifiers to predict Diabetes from different patients' information. In this dataset, the Logistic Regression classifier stands out with an accuracy of 77.21%

Table 3 Performance Comparison

Classifier Name	Correct Case	Incorrect Case	Accuracy
Bagging	582	186	75.78%
Logistic Regression	593	175	77.21%
Naive Bayes	586	182	76.30%
Random Forest	582	186	75.78%
Random Tree	523	245	68.05%

4) Hepatitis Disease: This disease is an inflammatory condition of the liver. Though this disease is an infection, various attributes of a patient need to be taken care of to determine the actual situation of the disease in a human body. In our applied data set, it has 155 instances with 19 various attributes. Linear Regression has better accuracy in a single ML approach with 85.16\%. The reason for the closest results among all single classifiers is the small number of instances.

The number of classes used in this data set result is two. They are: “DIE” and “LIVE”.

Table 4 Performance Comparison

Classifier Name	Correct Case	Incorrect Case	Accuracy
Bagging	126	29	81.29%
Logistic Regression	128	27	82.58%
Naive Bayes	131	24	84.52%
Random Forest	132	23	85.16%
Random Tree	119	36	76.77%

1.3 Deep Learning

Working with medical images is a comparatively new and widely appreciated concept because of accurate and specific results. Machine learning paves the way to understand medical images precisely and work with this vast field.

Throughout the years, two types of work are associated with a medical image and machine learning approach. One of them is to work on existing data and make them for a better result approach, and another one is predicting the disease and corresponding state by judging the parameters of a particular image. Predicting by analyzing the parameters of the image is our main concern.

We took a considerable number of instances from the data set and observed them with precise needs. We eliminated all the erroneous data because the model had a success rate of 89\% with missing and counterfeit data. Lastly, eliminate them by setting a threshold value for every parameter related to the medical image. After that, the prediction accuracy jumped up to 98.76%.

Nowadays, a lot of approaches are used for image detection and their application by ML classifiers. We will compare some of the fantastic working techniques and their accuracy, which will help portray our work accuracy regarding similar working principles.

Table 5 Performance Comparison

Author	Working Principle	Instance Number	Accuracy
Adam Hoover et al. [15]	Fuzzy Convergence	81	89%
Gagnon et al. [11]	Priori Geometric	40	90.2%
Hsiang Sing Naik [21]	SVM	591	96%
Proposed Method	CNN	1300	93.75%

The inception model of the Convolution Neural Network (CNN) is very effective in a vast number of samples. Our data set contains a massive number of instances, so the effective use of the inception model yields successful predictions with a higher rate of accuracy, which is 93.75%. On the other hand, all similar works have a smaller number of instances. By comparing their work with ours, we can say the inception model has better acceptability and productivity rate while working with the medical images.

Elimination of noisy data from the data set plays a vital role that helps to achieve better accurate results. This elimination should depend on the format of the image, redundancy, counterfeit image, etc.

2 CONCLUSION

Medical science is one of the most important aspects of any society. Medical privilege counts as the need for a human phenomenon. So an efficient approach and better accurate result can lead to a better situation in terms of medical science applications. Disease prediction and disease identification with computational assistance can help medical science to prosper a more significant concept. Machine Learning and Deep Learning are the two most convenient technologies for bettering disease prediction and identification.

In this paper, both shallow and deep learning were executed. The data was acquired from open sources and with a meaning formula. This method helped us to eliminate missing data, counterfeit symbols, and blank spaces. Data validation is also supported to get better data for execution. We then applied our five selective single classifier models. They are Bagging, Linear Regression, Bayes, Random Forest, and Random Tree. The experiment considered 1300 samples for experiment where almost 94 percent (93.75%) are identified correctly. In brief more, the number of instances yields better results.

References

1. World Health Organization. World Health statistics Annual (2006)
2. Bansal, P.: Intel Image Classification (2019)
3. Bhatla, N., Jyoti: An Analysis of Heart Disease Prediction using Different Data Mining Techniques. International Journal of Engineering Research & Technology (IJERT) **1**, 2278–0181 (2012)
4. Bouckaert, R.R.: (2013)
5. Cardiology, H.I.O., Janosi, M.D.B.A.: (1988), https://archive.ics.uci.edu/ml/datasets/Heart+Disease
6. Cardiology, H.I.O., Janosi, M.D.B.A.: (1988), https://archive.ics.uci.edu/ml/datasets/Hepatitis
7. Cardiology, H.I.O., Janosi, M.D.B.A.: (2019), https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29
8. Cardiology, H.I.O., Janosi, M.D.B.A.: (2019), https://archive.ics.uci.edu/ml/datasets/Diabetes
9. Cox, D.R.: Analysis of survival data. Chapman & Hall, London (1984)
10. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications **36**(4), 7675–7680 (2009). https://doi.org/10.1016/j.eswa.2008.09.013, https://dx.doi.org/10.1016/j.eswa.2008.09.013
11. Gagnon, L., Lalonde, M., Beaulieu, M., Boucher, M.C.: Procedure to detect anatomical structures in optical fundus images. Proc. SPIE Medical Imaging: Image Processing pp. 1218–1225 (2001)
12. Graf, A.B., Wichmann, F.A., Bülthoff, H.H., Schölkopf, B.: Classification of Faces in Man and Machine. Neural Computation **18**(1), 143–165

(2006). https://doi.org/10.1162/089976606774841611, https://dx.doi.org/10.1162/089976606774841611

13. Greenspan, H., Ginneken, B.V., Summers, R.M.: Guest Author Working Principle Instance Number Accuracy Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Transactions on Medical Imaging **35**(5), 1153–1159 (2016)
14. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R.: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA **316**(22), 2402–2402 (2016). https://doi.org/10.1001/jama.2016.17216, https://dx.doi.org/10.1001/jama.2016.17216
15. Hoover, A.D., Kouznetsova, .V., Goldbaum, .M.: (2000)
16. Hornik, K., Stinchcombe, M., White, H.: Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks **3**(5), 551–560 (1990). https://doi.org/10.1016/0893-6080(90)90005-6, https://dx.doi.org/10.1016/0893-6080(90)90005-6
17. Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective **23**(1), 89–109 (2001). https://doi.org/https://doi.org/10.1016/S0933-3657(01)00077-X, https://www.sciencedirect.com/science/article/pii/S093336570100077X
18. Kusiak, A., Dixon, B., Shah, S.: Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine **35**(4), 311–327 (2005). https://doi.org/10.1016/j.compbiomed.2004.02.004, https://dx.doi.org/10.1016/j.compbiomed.2004.02.004
19. Liao, S.C., Lee, I.N.: Appropriate medical data categorization for data mining classification techniques. Medical Informatics and the Internet in Medicine **27**(1), 59–67 (2002). https://doi.org/10.1080/14639230210153749, https://dx.doi.org/10.1080/14639230210153749
20. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005, https://dx.doi.org/10.1016/j.media.2017.07.005
21. Pace, J., Lee, N., Naik, H.S., Ganapathysubramanian, B., Lübberstedt, T.: Analysis of Maize (Zea mays L.) Seedling Roots with the High-Throughput Image Analysis Tool ARIA (Automatic Root Image Analysis). PLoS ONE **9**(9), e108255–e108255 (2014). https://doi.org/10.1371/journal.pone.0108255, https://dx.doi.org/10.1371/journal.pone.0108255
22. S, V., A, S.: Performance Comparison of OCR Tools (2015). https://doi.org/10.5121/iju.2015.6303, https://dx.doi.org/10.5121/

iju.2015.6303

23. Sandhya, N., M., M.: Analysis of Classification Techniques for Efficient Disease Prediction. International Journal of Computer Applications **155**(8), 20–24 (2016). https://doi.org/10.5120/ijca2016912388, https://dx.doi.org/10.5120/ijca2016912388
24. Sharma, K., McCue, P., Dunn, S.R.: Diabetic kidney disease in thedb/dbmouse. American Journal of Physiology-Renal Physiology **284**(6), F1138–F1144 (2003). https://doi.org/10.1152/ajprenal.00315.2002, https://dx.doi.org/10.1152/ajprenal.00315.2002
25. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M.: Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Transactions on Medical Imaging **35**(5), 1196–1206 (2016). https://doi.org/10.1109/tmi.2016.2525803, https://dx.doi.org/10.1109/tmi.2016.2525803
26. Team, S.: A List of Technology Abbreviations Commonly Used (12 2018), https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-summary/
27. Temurtas, H., Yumusak, N., Temurtas, F.: A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications **36**(4), 8610–8615 (2009). https://doi.org/10.1016/j.eswa.2008.10.032, https://dx.doi.org/10.1016/j.eswa.2008.10.032
28. Vapnik, K.: Statistical learning theory. Wiley, New York (1998)
29. Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C.: High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. NeuroImage **50**(4), 1519–1535 (2010). https://doi.org/10.1016/j.neuroimage.2009.12.092, https://dx.doi.org/10.1016/j.neuroimage.2009.12.092
30. Wernick, M., Yang, Y., Brankov, J., Yourganov, G., Strother, S.: Machine Learning in Medical Imaging. IEEE Signal Processing Magazine **27**(4), 25–38 (2010). https://doi.org/10.1109/msp.2010.936730, https://dx.doi.org/10.1109/msp.2010.936730
31. Xing, Y., Wang, J., Zhao, Z., Gao, Y.: Com- bination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. International Conference on Convergence Information Technology (2007)
32. Yan, K.: (07 2018), https://nihcc.app.box.com/v/DeepLesion