International Journal of Advanced Trends in Engineering and Management
(ISSN 2583-7052)

Vol. V, No. 06, pp. 34-42, June

Improved ML Algorithms for Medical Image Analysis, Disease Diagnosis and Patient Risk Prediction in Health Care

Pradeep KGM[1], Dasari Rammurthy[2], Mathesh G[3]

[1] Associate Professor & HOD, Department of Electrical and Electronics Engineering,
Rajeev Institute of Technology, Hassan, Karnataka, India.
eehod@rithassan.ac.in
[2] Professor & Dean (Academics), Department of Electronics and Communication Engineering,
Rajeev Institute of Technology, Hassan, Karnataka, India.
deanacademics@rithassan.ac.in
[3] Assistant Professor, Department of Electrical and Electronics Engineering,
Rajeev Institute of Technology, Hassan, Karnataka, India.
matheshg@rithassan.ac.in

ABSTRACT: In the field of healthcare, data mining techniques hold significant promise for enhancing patient care, disease prevention, and healthcare management. Complex data mining methods, like deep learning, can lead to overfitting and produce difficult-to-interpret black-box models. The most challenging Android assistance application to locate nearby hospitals emerges as a crucial, potentially life-saving endeavour and many more features are included. The MIMIC-III dataset offers a rich source of de-identified health data from Beth Israel Deaconess Hospital in Boston, Massachusetts, encompassing electronic health records, vital signs, lab results, and more. Quality-based images and some DL techniques are utilized for this kind of application. The process involves collecting and preprocessing medical images, including segmentation to isolate regions of interest. Regarding RP methodologies, RF stands out as a versatile and robust option, capable of handling diverse data types while mitigating overfitting risks. While SVM, GBM, and NN also have their strengths, RF's interpretability, ease of training, and reliability make them suitable for various risk prediction tasks. Appropriate algorithms are applied for analysis, and performance is evaluated using validation techniques.

Keywords: Machine Learning, Risk Prediction, Image Analysis, Medical Diagnosis, Correlation

1. Introduction

The need for medical image analysis is driven by several crucial factors in the field of healthcare. It plays a pivotal role in disease diagnosis, treatment planning, and patient care. Timely detection of diseases is a fundamental aspect of healthcare. By pinpointing abnormalities at an early stage, it empowers healthcare professionals to initiate treatment promptly, often improving patient outcomes and reducing healthcare costs. Medical images provide a wealth of information that is sometimes beyond the human eye's capacity to discern. In planning treatment strategies, medical image analysis plays a pivotal role. It enables healthcare providers to visualize the affected areas in three dimensions, aiding in surgical planning, radiation therapy, and the placement of medical implants. For chronic diseases and conditions that require long-term management, regular monitoring is essential. The era of personalized medicine relies heavily on medical image analysis. By automating repetitive tasks, medical image analysis increases efficiency in healthcare settings.

Assessing risk prediction models is a crucial step in evaluating their performance and reliability.

Begin with an introduction that defines the context. Discuss the significance of accurate risk assessment. Highlight how it can inform decision-making, reduce uncertainty, and lead to better outcomes. Provide a clear definition of risk prediction models. Describe different types of risk prediction models. Accurate risk prediction depends on the quality and quantity of data. Describe how data is collected, cleaned, and preprocessed to ensure its suitability for modelling. The process of developing risk prediction models. The importance of cross-validation to prevent overfitting. Emphasize the need for external validation on independent datasets to assess a model's generalizability and robustness. Compare the performance of your risk prediction model with baseline models or random chance to demonstrate its effectiveness.

Performance correlation is used to identify similarities or patterns within images. Correlation is often applied in tasks such as template matching, feature detection, and object recognition. Start by importing the required libraries and packages for image processing and correlation. Load the two images which has to be correlated. These images should have the same dimensions, and they can be in grayscale or colour. Convert them to the appropriate data format for the chosen library. Depending on your specific application, it might need to pre-process the images. Use the correlation function provided by your chosen library to compute the correlation between the two images. The result of the correlation will be a 2D array, where each pixel value represents the correlation coefficient at that position. Higher values indicate a stronger correlation between the two images. Depending on your application, it may want to apply a threshold to the correlation result to identify regions or points where the correlation is above a certain threshold.

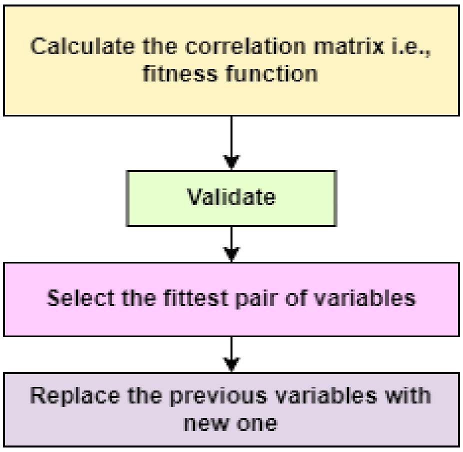

Figure 1: *Performing Correlation Analysis*

2. Recent Works

Senerath Mudalige Don Alexis Chinthaka Jayatilake et al [1] has examined on multiple methodologies related to ML and DL. Effective ML algorithm selection relies on classifier accuracy and log loss. Clustering algorithms require careful parameter tuning, especially for high-dimensional, non-spherical biological and biomedical data. Identifying the best machine learning algorithm depends on various factors like dataset characteristics, pre-processing, selection of features, & problem domain. It's an iterative process based on experience and domain knowledge. Algorithm performance varies across different healthcare domains. KNN is unsuitable for cancer predictions due to complex and large datasets. DNN demonstrate better performance for breast cancer and diabetes prediction. ML aids clinicians by highlighting critical factors for treatment decisions and assessing disease severity based on patient data. Integrating medical imaging with 3D printing enhances surgical planning, leading to precise patient-specific interventions.

Christian Tchito Tchapga et al [2] has introduced a Spark CNN method for easy detection of diseases related to any kind. The primary objective of medical image classification is to identify disease-infected parts of the human body while achieving high accuracy. The classification process involves two main steps building a classifier model and using the model to predict the class of unlabeled biomedical images. The training phase includes data presentation from the labelled biomedical image dataset, feature extraction, and model training using techniques like GDA. Labelled images, part of the training dataset, are crucial for supervised learning to describe class/group characteristics. Feature extraction involves identifying relevant image descriptors and creating feature vectors, particularly deep features in DL. SVM and CNN are explained as ML algorithms used for classification, with SVM being effective for medium-sized datasets. In the testing phase, unlabelled biomedical image feature vectors are classified based on the previously trained model, involving capturing, feature extraction, model application, and prediction.

V. Durga Prasad Jasti et al [3] has proposed a LS-SVM for the early detection of diseases based on the medical images. MIAS and DDSM databases are mentioned as commonly used databases in breast cancer research, with a focus on analyzing the MIAS database in this project. Image preprocessing is conducted using a geometric mean filter to remove various types of noise present in mammogram images, which is crucial for accurate disease classification. Features are extracted using AlexNet, a DL approach with 22 layers of feature extraction based on transfer learning, plus a fully connected layer. The relief algorithm, inspired by instance-based learning techniques, assigns proxy statistics to each feature to determine its relevance to the target concept, aiding in feature selection. LS-SVM is highlighted as a more cost-effective and user-friendly alternative to SVM, particularly for linear equations.

Tasmia Tahmida Jidney et al [4] has focused on AutoML methodology for the detection of diseases. MIA involves the analysis of medical images, initially stemming from Computer Vision, and has become vital in modern medicine for non-invasive examinations. This involves enhancing raw medical image data for visualization and analysis, utilizing edge detection and region-growing techniques. Initially focused on rule-based systems, medical image analysis now incorporates advanced machine learning techniques like deep learning to automate and improve accuracy. AI algorithms, trained on annotated medical images like MRI scans, can accurately detect brain tumors, though human validation is necessary. AI has been integrated into medical imaging systems, enhancing diagnostic accuracy, particularly in early disease detection. Automated machine learning (AutoML) simplifies and accelerates AI model development, making AI more accessible and efficient, especially in regulated industries like healthcare. AutoML includes automated feature engineering and hyperparameter optimization, improving model precision and efficiency.

Nafiseh Ghaffar Nia et al [5] has proposed a AI and DL methodologies with multiple enhancement of diseases and predictions. AI technologies are increasingly being employed in medicine to improve disease diagnosis, prediction, and treatment. Learning algorithms and big data from medical records and wearables are vital for effective AI implementation in healthcare. AI models enhance disease diagnosis, classification, and treatment selection, leading to better patient outcomes. It can detect dangerous tumors in medical images, enabling early diagnosis and treatment. This models effectively identify undiagnosed or less-diagnosed patients, including those with rare diseases. ML & DL techniques are increasingly used to diagnose heart diseases, including coronary atherosclerotic heart disease. These methods, such as SVM, ANN, and Decision Trees, achieve high accuracy in diagnosing breast cancer, with accuracy rates exceeding 95%. These are applied to predict and categorize genetic disorders, although challenges remain due to diverse genotypes. This technology is employed for the timely identification of neurological disorders such as Alzheimer's and Parkinson's disease, as well as skin cancer. CNN-based models are prevalent in this context.

Table 1: *Matriculation Analysis on Existing Approaches*

Author	Algorithm	Merits	Demerits	Accuracy
Senerath Mudalige Don Alexis Chinthaka Jayatilake et al	ML, DL.	Analysis was performed on multiple diseases.	Only the theoretical discussion was performed.	75%
Christian Tchito Tchapga et al	Spark, CNN	Limited layers and accurate prediction.	Only for medium datasets the performances is high.	97.34%
V Durga Prasad Jasti et al	LS-SVM	For every process a specific techniques was utilized.	Only particular disease was identified.	98%
Tasmia Tahmida Jidney et al	Auto ML	The treatment identification was efficient.	It has to improve the integration with other technologies.	88.2%
Nafiseh Ghaffar Nia et al	DL, ML, AI	A clear report about each method was described.	Hybrid methods are not discussed.	95%

3. Proposed Work Explanation

Healthcare detection using data mining approaches has shown great promise in improving patient care, disease prevention, and healthcare management. The quality and accuracy of healthcare records are very important to data mining. Issues including noisy data, missing values, and data input mistakes plague a lot of healthcare datasets. There are problems in a lot of healthcare records, such as missing values, noise data, and mistakes in data entry. Most of the time, the data are saved in different methods and forms. Overfitting happens when a DM methodology is extremely complicated & exactly fits the trained data with too close. Many advanced data mining techniques, such as deep learning, create complex, black-box models that are challenging to interpret. Data mining approaches are typically constrained by the data available. They may not capture all relevant factors influencing a patient's health.

Developing an Android assistance application to help individuals find the nearest hospitals is a valuable and potentially life-saving endeavor. An Android assistance application designed to help individuals find the nearest hospitals can serve as a vital tool in such situations. The app uses GPS technology to determine the user's current location and provides a list of hospitals in proximity. Seamless integration with navigation apps like Google Maps allows users to get directions to the selected hospital. The app provides real-time information on the availability of beds, emergency room wait times, and current patient loads to help users make informed decisions. An integrated medication tracker helps users manage their medications, set reminders for doses, and maintain a medication history. Users can make informed decisions about where to seek medical help based on real-time data, reviews, and hospital specialties. It promotes emergency preparedness by offering guidance on what to do in case of various medical emergencies.

Collect the medical images you need for analysis. Clean and pre-process the images to improve their quality and consistency. If needed, perform image segmentation to identify and isolate regions of interest within the medical images. This can involve techniques like thresholding, region growing, or deep learning-based methods. From the divided areas, extract pertinent characteristics. These traits might be related to a particular domain, such as form, texture, quantity, or other attributes. Apply appropriate algorithms or models to analyze the extracted features. This could involve machine learning, deep learning, or traditional statistical methods for tasks such as classification, detection, or regression. Evaluate the performance of your analysis using appropriate validation techniques. The pre-processing algorithm is discussed below as flow chart

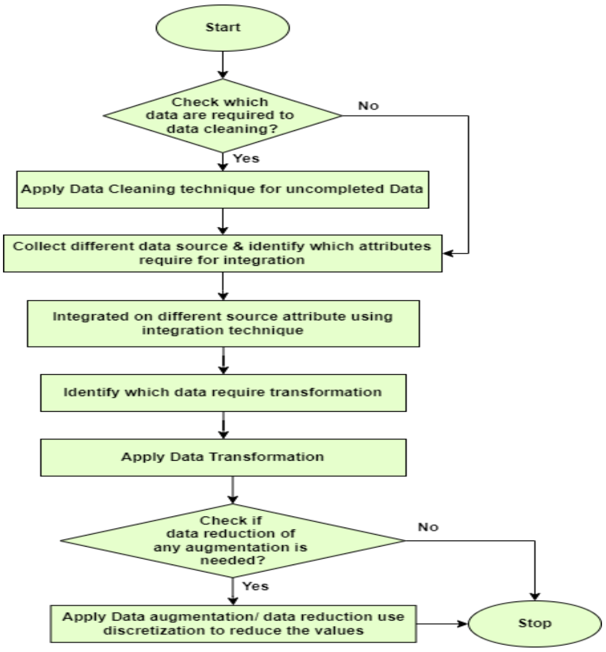

Figure 2: *Flow Chart for Pre-processing*

The best algorithm to perform risk prediction depends on the specific problem and the available data. RF are ensembles of decision trees that are trained on different subsets of the data. This makes them more robust to overfitting and able to handle complex relationships between features. GBMs are also ensembles of decision trees, but they are trained sequentially. This allows them to learn from the mistakes of previous trees and improve their performance over time. SVM findings a hyperplane in the data that separates the positive examples from the negative examples with the largest possible margin. NN can be very effective for risk prediction tasks, but they can also be computationally expensive to train.

RF is a good all-around algorithm for risk prediction. They are relatively easy to interpret and train, and they can handle a variety of different data types. They are also less likely to overfit than other algorithms, which makes them more reliable for real-world applications. To make a prediction, the random forest model inputs the data into each of the decision trees and then averages the predictions. The average prediction is then used as the final prediction of the model. They are accurate and robust to overfitting. They can handle a variety of different data types. The interpretation and training of these entities are quite straightforward. These tools have the capability to forecast a diverse array of dangers.

The working of Random Forest algorithm is shown below

Input: Symptoms Dataset

Output: Prediction of Disease

Step 1: Determine the target variable's mean or average.

Step 2: Compute each sample's residuals in step two.

Step 3: Create a decision tree in step three. The objective is to anticipate the Residuals by constructing a tree.

Step 4: Use every tree in the ensemble to predict the target label.

Step 5: Determine the updated residuals

Step 6: Repetition of steps 3 through 5 until the hyper parameter's (numbers of estimators) set number of iterations is reached.

Step 7: After training, determine the target variable's value by using all of the trees in the ensemble. The mean which is calculated in Step 1 plus all of the residuals projected by the forest's ind ividual trees compounded by the learning rate will equal the final forecast.

Medical Diagnosis involves a series of steps designed to gather information, assess symptoms, perform tests, and arrive at a conclusive diagnosis. Based on the history and physical examination, healthcare providers may order a battery of laboratory tests and diagnostic procedures. Once the initial test results are available, healthcare professionals create a list of potential diagnoses, known as the differential diagnosis. In some cases, confirmation of the diagnosis requires additional testing, such as a biopsy or a second opinion from a specialist. With a confirmed diagnosis, healthcare experts create a treatment plan that is customised to the patient's situation. It is crucial to educate the patient about their diagnosis, treatment options, potential risks, benefits, and expected outcomes. Collaboration among healthcare professionals, ongoing monitoring, and patient education are integral components of this process, ensuring the best possible care and outcomes for patients.

The algorithm for medical diagnosis is shown below.

```
Input: Dataset of diseases
Output: Medical Diagnosis
df = create_dataframe(data)
cleaned_data = preprocess_data(df)
X, y = split_features_and_target(cleaned_data)
X_train, X_test, y_train, y_test = split_train_test_data(X, y)
selected_model = choose_model()
train_model(selected_model, X_train, y_train)
y_pred = evaluate_model(selected_model, X_test)
accuracy = calculate_accuracy(y_test, y_pred)
report = generate_classification_report(y_test, y_pred)
display_results(accuracy, report)
interpretability_methods = choose_interpretability_methods()
apply_interpretability_methods(selected_model, X_test, interpretability_methods)
ensure_ethical_compliance()
ensure_regulatory_compliance()
deploy_model()
while continuous_monitoring:
   monitor_model_performance()
   if performance_degraded:
      retrain_model()
```

Disease diagnosis encompasses various methods and approaches used by healthcare professionals to identify and determine the presence of diseases or medical conditions in patients. Clinical diagnosis relies on a healthcare provider's expertise in evaluating a patient's signs and symptoms. Laboratory diagnosis involves the analysis of bodily fluids, tissues, and cells to identify diseases. X-rays, CT scans, MRI, and ultrasound are all forms of imaging are used to visualize internal structures of the body. Endoscopy involves the use of specialized instruments with cameras to examine the interior of organs or body cavities. A biopsy involves the removal of a small tissue sample from the body for microscopic examination. Genetic testing examines an individual's DNA for mutations or variations associated with genetic disorders, susceptibility to diseases, or pharmacogenetics.

Extracting information from X-rays using neural networks is a common and valuable application of artificial intelligence in the field of healthcare and medical imaging. Convolutional neural networks (CNNs) in particular have demonstrated considerable potential in tasks including object identification, segmentation, X-ray image classification, and illness diagnosis. Compile a collection of X-ray pictures. To increase the resilience of the network, preprocess the photos by resizing, normalizing, and augmenting the images. Utilizing the training set, train the neural network. The loss function and optimization algorithm may vary depending on the specific task. Neural networks can process X-ray images rapidly, reducing the time required for diagnosis and information extraction. This is particularly crucial in emergency situations where quick decisions are necessary. Neural networks can provide highly accurate results in X-ray analysis tasks.

4. Results and Discussion

Materials Available: MIMIC-III is a comprehensive dataset containing de-identified health data from The individuals that were taken in to the Beth Israel Deaconess Hospital are located in Boston, Massachusetts. The dataset encompasses a diverse array of clinical information such as EHRs, vital signs, lab results, and more. The dataset is available through the PhysioNet platform. Researchers need to request access and adhere to specific data usage agreements. The ChestX-ray14 dataset is a collection of chest X-ray images along with text-mined labels for 14 different thoracic pathologies. It was previously available on the NIH website, but it should verify its current availability and terms of use. This dataset comprises over 100,000 chest X-ray images annotated with various thoracic pathologies. It's intended for research in computer-aided diagnosis and image analysis. The dataset was available from the NIH, but access details and usage terms may have changed, so please verify the current status.

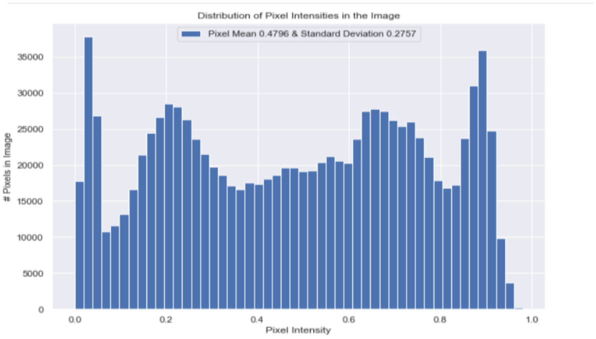

Figure 3: *Variation in Pixel Intensity based on Distribution of Image*

The distribution of pixel intensities in an image can provide valuable information about the image's content, contrast, and other characteristics. To visualize the distribution of pixel intensities in an image, model can create a histogram as shown in figure 3

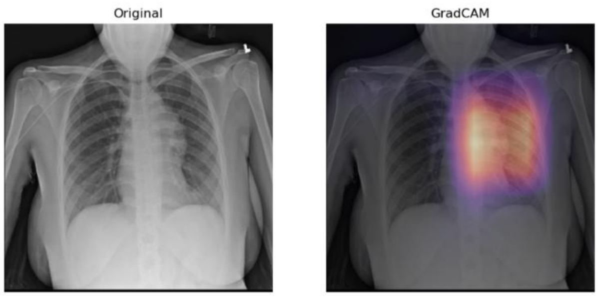

Figure 4: *GradCAM Converted image*

Figure 4 presents a Grad-CAM Image which is a technique used for visualizing and interpreting the decisions made by a convolutional neural network (CNN) in the context of image classification tasks. It helps you understand which regions of an image were influential in the network's prediction by highlighting the areas that had the most impact on the final decision.

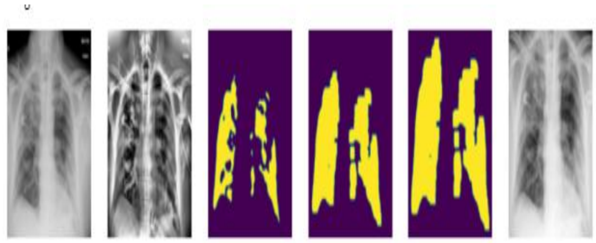

Figure 5: *UNET Segmentation Approach*

The top part of the U-Net, known as the encoder, is a series of convolutional layers that gradually reduce the spatial dimensions while increasing the number of feature channels. This part extracts features from the input image and is presented in figure 5.

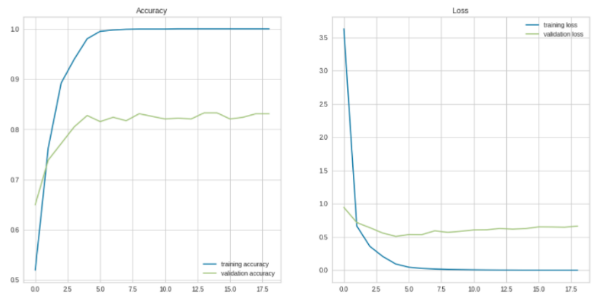

Figure 6: *Accuracy & Validation of Proposed Model*

Model would typically monitor both accuracy and loss during the training process. A decrease in the loss indicates that the model is learning, while an increase in accuracy on the validation dataset means that it is making better predictions. Figure 6 represents both the measurements.

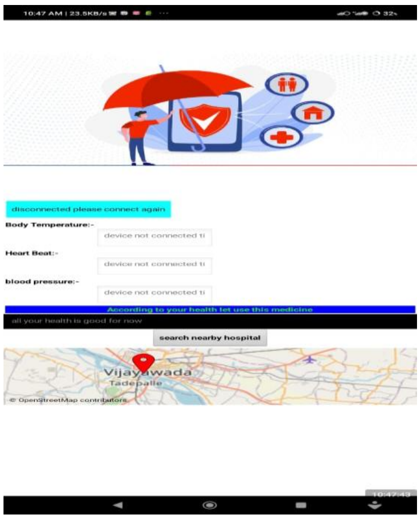

Figure 7: *Measuring Human Body Parameters*

For all three vital signs (body temperature, heart rate, and blood pressure), the data collected by IoT sensors can be securely transmitted and stored. This data can be accessed by individuals for personal health management or by healthcare professionals for remote patient monitoring.

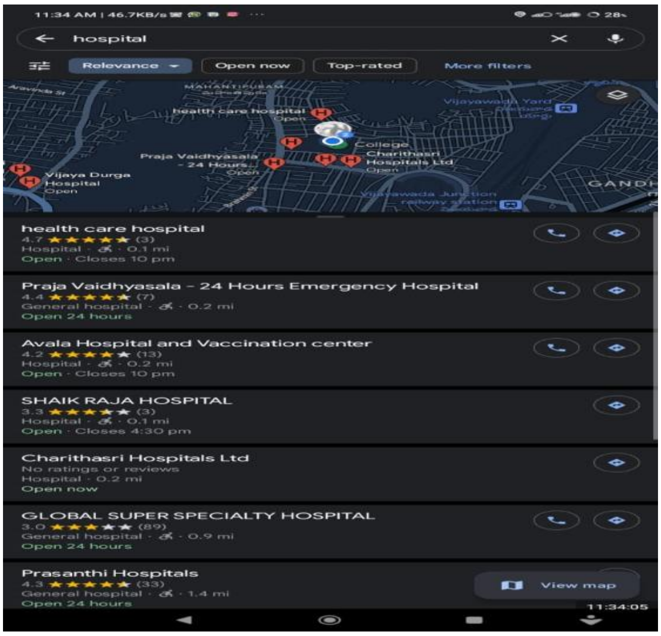

Figure 8: *Finding Nearest Hospitals*

An algorithm is designed to calculate the distance between the user's location and the locations of nearby hospitals. Various distance measurement algorithms can be used, such as the Haversine formula for calculating distances between two points on the Earth's surface.

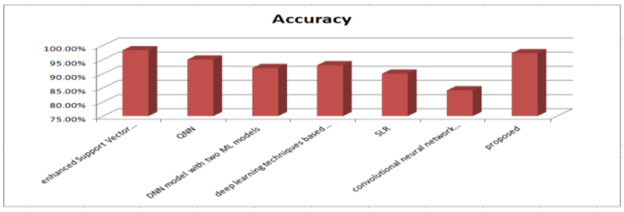

Figure 9: *Accuracy Analysis*

Accuracy analysis should be an ongoing process, with regular assessments and improvements to ensure that the IoT-based healthcare system provides reliable and accurate information and services to users and healthcare providers. Out of all the approaches, the proposed model has highest accuracy.

5. Conclusion

In the realm of medical diagnosis, a comprehensive process involves gathering patient information, symptom assessment, testing, and arriving at a conclusive diagnosis. Collaboration among healthcare professionals, patient education, and ongoing monitoring play pivotal roles in delivering optimal care and outcomes. The MIMIC-III dataset offers a rich source from different areas of the medical field and contains information about patients. RF averages predictions from multiple decision trees for improved accuracy. While SVM, GBM, & NN also have their strengths, RF's interpretability, ease of training, and reliability make them suitable for various risk prediction tasks. Relevant features are extracted from these areas, which can be related to form, texture, or other attributes. Various methods for disease diagnosis include clinical evaluation, laboratory testing, imaging techniques, endoscopy, biopsies, and genetic testing. Each method contributes to the accurate identification of diseases and medical conditions.

References

1. Senerath Mudalige Don Alexis Chinthaka Jayatilake; Gamage Upeksha Ganegoda, Year: 2021, “Involvement of Machine Learning Tools in Healthcare Decision Making”, In Journal of Healthcare Engineering, Vol: 2021. Hindawi Limited.
2. Christian Tchito Tchapga; Thomas Attia Mih; Aurelle Tchagna Kouanou; Theophile Fozin Fonzin; Platini Kuetche Fogang; Brice Anicet Mezatio; Daniel Tchiotsop, Year: 2021, “Biomedical Image Classification in a Big Data Architecture Using Machine Learning Algorithms”, In Journal of Healthcare Engineering, Vol: 2021. Hindawi Limited.
3. V. Durga Prasad Jasti; Abu Sarwar Zamani; K. Arumugam; Mohd Naved; Harikumar Pallathadka;

F. Sammy; Abhishek Raghuvanshi; Karthikeyan Kaliyaperumal, Year: 2022, “Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis,” Security and Communication Networks, 2022.

4. Tasmia Tahmida Jidney; Angona Biswas; Nasim Md Abdullah Al; Ismail Hossain; Md Jahangir Alam; Sajedul Talukder; Mofazzal Hossain; Md Azim Ullah, Year: 2023, “AutoML Systems For Medical Imaging”.
5. Nafiseh Ghaffar Nia; Erkan Kaplanoglu; Ahad Nasab, Year: 2023, “Evaluation of artificial intelligence techniques in disease diagnosis and prediction”, Discover Artificial Intelligence, Vol: 3, No: 1.