International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Machine Learning for Hospital Health Prognostication: Next Gen Health Monitoring

Ayesha Siddiqa[1], Kavana G S[2], Manish S Patel[3], Lisha M[4], Madhushree J S[5]

[1]Assistant Professor, Department of CSE JNN College of Engineering.
[2] Department of CSE JNN College of Engineering.
[3] Department of CSE JNN College of Engineering.
[4] Department of CSE JNN College of Engineering.
[5] Department of CSE JNN College of Engineering.

Abstract - The diagnosis of critical diseases like breast cancer, diabetes, malaria, and pneumonia is challenging, especially in under-resourced healthcare settings due to infrastructure limitations and time-consuming methods. To address these issues, we propose a web-based Multiple Disease Predictor platform, a centralized diagnostic tool powered by machine learning algorithms. The platform processes medical images, blood test results, and clinical data to deliver accurate, real-time disease predictions. By consolidating multiple diagnostic capabilities into one interface, the system reduces reliance on specialized equipment, improves efficiency, and enhances diagnostic accuracy, particularly in resource-limited areas. Advanced machine learning and deep learning techniques enable rapid, accurate data analysis, minimizing diagnostic errors and misdiagnosis risks, which is crucial in underserved regions with limited access to specialized equipment. The platform also supports early detection and timely treatment, improving patient outcomes. By integrating diverse diagnostic data sources, healthcare providers can make faster, informed decisions. This system democratizes access to advanced diagnostic technology, promoting cost-effective healthcare.

Key Words: Diagnosis, Diabetes, Machine Learning, Decision tree, Logistic Regression and Accuracy.

1. INTRODUCTION

The Multiple Disease Prediction Project leverages machine learning and data analytics to predict critical diseases such as diabetes, breast cancer, malaria, and pneumonia, aiming to enhance early detection and improve patient outcomes, especially in resource-limited settings. By integrating various diagnostic tools into a single platform, the project simplifies the diagnostic process, allowing healthcare professionals to predict multiple diseases from one interface, saving time and reducing errors.

The project uses predictive models to analyse patient data, such as glucose levels for diabetes, mammograms and histopathological images for breast cancer, blood smear images for malaria, and chest X-rays for pneumonia. Machine learning techniques like logistic regression, decision trees, and Convolutional Neural Networks (CNNs) are employed to accurately classify patients and detect early signs of these diseases. For example, CNNs have proven highly effective in detecting malignancies in breast cancer and classifying malaria-infected cells in blood smear images. The use of these advanced machine learning techniques not only improves diagnostic accuracy but also accelerates the detection process, ensuring timely intervention and better patient outcomes.

2. LITERATURE REVIEW

Yasodhaet al. [1] uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared.

The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others. Temperature, precipitation and humidity influence the mosquito's life cycle for their growth and development [2].

The vector larvae survive only when the environmental condition is conducive in a moderate temperature above 160C and die when it is lower or higher [3]. The rate at which mosquitoes bite humans by sucking their blood increases when the environmental conditions are favorable for their survival.

Fatima B., et d in [4] define an uncertain expert system for breast cancer forecast to additional support the procedure of breast cancer analysis. This method is accomplished enough to capture vague and imprecise information prevalent in classification of breast cancer. For this the paper utilized a uncertain reasoning model, which has high interpretability early diagnose of the accuracy of the system with an average 95% which shows the advantage of the system in the forecast process compared to other related work. Breast cancer analysis and forecast were two medical requests, which position as great test to the investigates. Machine learning and data mining methods usage has transformed the entire practice of breast cancer Diagnose and Forecast. Breast cancer Diagnose decides

design from breast lump and breast cancer Diagnose and Forecast. Breast Cancer Forecast predicts while Breast Cancer is probable to return in patients that had their cancers removed. Thus, these two problems were mainly in the scope of the organization problems. This study paper encapsulates various reviews, technical articles on breast cancer diagnosis & prognosis.

P. Pratik, Hemprasad Patil, [5] Early detection and prompt treatment can significantly reduce the mortality rate associated with this disease. Chest x-rays are commonly employed to identify the symptoms of the disease. In this study, a CNN-based model was utilized to automatically detect crucial features, as opposed to using hand-engineered feature selection techniques.

Gupta et al. [6] aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyses the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes RapidMiner and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows highest accuracy i.e. 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and RapidMiner. after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in a dataset having dichotomous values, which means that the class variable has two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model.

Sri Hari Nallamala, et al. [7] describes an assortment of web use mining practice can propel exertion on various regions of logical, restorative and online networking applications to progress toward for the exploration and security joined zone.

Vazirani, at all [8] suggests the 2 NN models, BPNN (Back Propagation Neural Network) and RBFN (Radial Basis Function). The expansion is finished utilizing a probabilistic total guideline. Presently, the measured neural system gave a precision of 95.75% over preparing information and 95.22% over testing information, which was tentatively resolved to be superior to solid neural systems.

Karabatak M., et al [9] suggests an automatic diagnosis scheme for detection breast cancer grounded on AR (Association Rules) and NN (Neural Network). Here, AR is used for sinking the measurement of intelligent classification. The projected AR+NN (combining 2 approaches) scheme performance is contrasted with NN model. The length of input feature space is condensed from nine to four by using AR. In test phase, 3–fold cross validation approach is applied on the WBC database to assess the projected system is 95.6%. This researcher established the AR cab be used for plummeting the length of feature and proposed AR+NN model can be used to discover rapid automatic diagnosed system for extra diseases. In Retrieval Number: B12600782S319/19©BEIESP. restorative areas where information and examination driven research are decidedly connected, new and unique research bearings were perceived to additionally propel the facility and natural.

3. PROPOSED METHEDOLOGY

Model Selection and Comparison:
Next, different training models will be selected and trained on the preprocessed dataset. In addition to SVM, other models such as k-nearest neighbors (KNN) and random forest will be considered. Each model will be evaluated using appropriate metrics like accuracy, precision, recall, and F1 score. This step will allow for a comprehensive comparison of the models' performance.

Data Handling and Filtering:
The first step in the project implementation is to handle and filter the data using the panda's library. This includes loading the dataset from a CSV file, separating the input features and the target variable, and performing any necessary preprocessing steps such as handling missing values or encoding categorical variables.

Model Selection and Comparison:
Next, different training models will be selected and trained on the preprocessed dataset. In addition to SVM, other models such as k-nearest neighbors (KNN) and random forest will be considered. Each model will be evaluated using appropriate metrics like accuracy, precision, recall, and F1
score. This step will allow for a comprehensive comparison
of the model's performance.

SVM Model Training:
Based on the comparison results, the SVM model, which achieved the highest accuracy of 98.8%, will be selected for further implementation. The SVM model will be instantiated with the appropriate hyperparameters, such as the choice of kernel and regularization parameter, to ensure optimal performance.

NumPy:
NumPy is a broadly useful cluster handling bundle which gives an elite multidimensional exhibit item and apparatuses for working with these exhibits. NumPy is the crucial bundle for logical processing with python.

Pandas:
A panda is an open-source Python Library giving superior information control and examination apparatus utilizing its amazing information structures. For information robbing and readiness, Python was significantly utilized. It had next to no commitment towards information examination. Pandas tackled this issue. Utilizing this, five run of the mill ventures Retrieval Number: B12600782S319/19©BEIESP.

Matplotlib:
Matplotlib is a Python 2Dplotting library which produces production quality figures in an assortment of printed version groups and intuitive conditions crosswise over stages. Matplotlib can use in python contents, the python and python shell, the Jupyter note pad, web application server, and four graphical UI toolbox. Matplot attempts to make simple thongs simple things simple and hard things conceivable. You can create plots, histogram, powers spectra, bar outlines, mistake

diagrams, disperse plots, and so forth, with only a couple of lines code.
Scikit-learn:
Scikit-learn gives a scope of managed and unsupervised learning calculations by means of a predictable interface in Python. It is authorized under a lenient rearranged BCD Licensed and id appropriated under numerous Linux disseminations, empowering scholarly and business use. The library is based upon the SciPy (logical python) that must be introduced before you use scikit-learn. This stack incorporates: NumPy: Base n-dimensional array package. Pandas: Data structures and analysis Matplotlib: Comprehensive 2D/3D plotting. I Python: Enhanced interactive console SciPy: Fundamental library for scientific computing. Sympy: Symbolic mathematics.

The methodology for disease prediction using machine learning typically begins with clearly defining the problem, which includes identifying the target disease and specifying the objectives, such as early detection, prognosis, or risk classification. This is followed by extensive data collection from credible sources like hospitals, clinical trials, or publicly available health datasets, ensuring the data encompasses all relevant attributes such as patient demographics, symptoms, lifestyle factors, lab test results, and imaging data if applicable. Ensuring the ethical use of data, such as anonymization and compliance with regulations like GDPR, is a critical step during this phase.

The collected data is then pre-processed to handle missing or inconsistent values, normalize numerical features, and encode categorical variables to make the dataset machine-readable. Data preprocessing also involves identifying and mitigating outliers, scaling features, and splitting the dataset into training, validation, and testing sets. Feature selection and engineering play a crucial role here, as domain knowledge and statistical techniques are used to identify or create the most predictive variables, ensuring the model focuses on meaningful inputs.
Multiple machine learning algorithms are then applied to the dataset, such as logistic regression, random forests, support vector machines, or deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for complex data like medical images or time series.

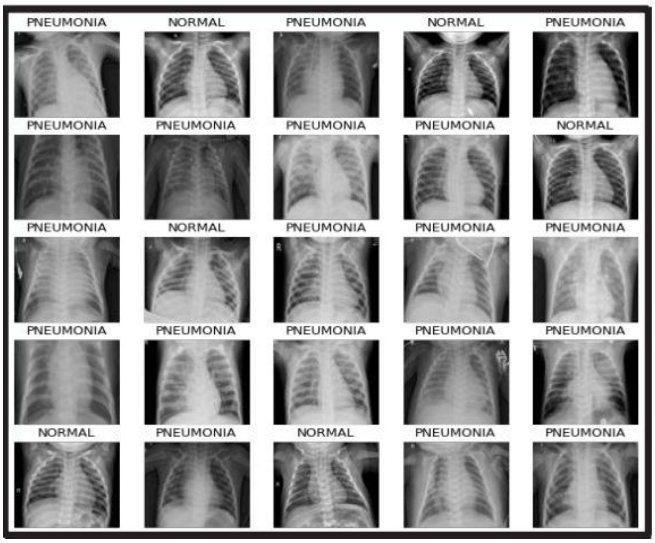

Fig -1: The image contains the dataset containing images with a label (pneumonia or normal)
The data set used in this work contains 5,863 chest x-ray images (from Kaggle). This dataset is further divided into test, train and validation images. These images in the dataset
are further divided into two categories, they are namely normal and pneumonia infected images.

The models are evaluated using performance metrics such as accuracy, sensitivity, specificity, F1-score, and AUC-ROC, ensuring they perform well in identifying true positives and minimizing false negatives, which is particularly critical in disease prediction scenarios. The best-performing model is then tested on unseen data to verify its generalization capability. Once validated, the model is deployed as part of a decision-support system for clinicians or integrated into patient monitoring applications. Post-deployment, the model undergoes regular updates and monitoring to incorporate new data and adapt to changing medical knowledge, ensuring consistent and reliable predictions over time.

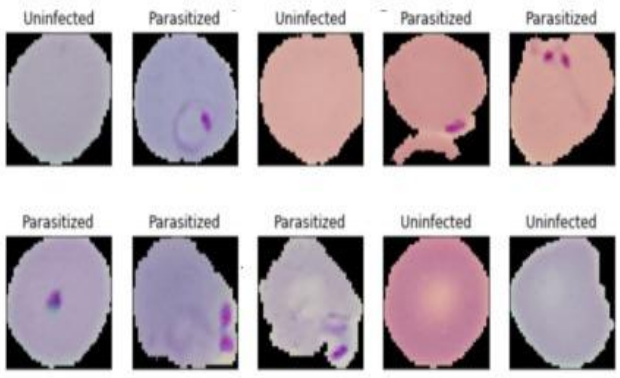

Fig -2: Example of malaria images to be used during the training and testing process.

The image shows blood smear samples, where cells are labelled as either **Parasitized** or **Uninfected**. These labels are commonly used in malaria detection to distinguish between healthy blood cells and those infected by Plasmodium parasites.

4. SYSTEM ARCHITECTURE

We have mixed structured and unstructured data in the healthcare fields to determine disease risk in this project. The use of a latent factor model to recreate missing data in medical records obtained from online sources. We could also assess the major chronic diseases in a specific area and population using statistical information. We consult hospital experts to learn about useful features when dealing with structured data. In the case of unstructured text files, we use the random forest algorithm to automatically select features.

A. Data collection:

Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered. The symptoms of the disease are collected from different health related websites.

Data Preprocessing:

Before feeding the data into the Prediction model, following data cleaning and preprocessing steps are performed

- Checking null values and filling using forward fill method
- Converting data into different cases
- Standardizing the data using mean and standard deviation
- Splitting the dataset into training and testing sets

B. Building Model:

Many methods are used to perform data mining. Machine learning is one of the approaches. Random forest Machine learning strategies include grouping, clustering, summarization, and many others. Since classification techniques are used in this project, classification is one of the data mining processes in this phase of categorical data classification. And this step is divided into two phases: training and testing. In the training phase, predetermined data and associated class labels are used for classification. The training stage is often referred to as supervised learning. The preparation and testing phases of the classification process are depicted in the diagram. In the training process, training tuples are used, and in the test data phase, test data tuples are used, and the classification rule's accuracy is calculated. Assume that the classification rule's accuracy on testing data is sufficient for the rule to be used for classification of unmined data.

C. Prediction:

Prediction using Random Forest:
Prediction done by Random Forest Model using Flask frame work model trained by training chronic disease dataset.

5. RESULTS

Table 1: - shows the accuracy achieved using random forest for each disease.

Diabetes	Machine Learning Model	98.25%
Breast Cancer	Machine Learning Model	98.25%
Malaria	Deep Learning Model (CNN)	96%
Pneumonia	Deep Learning Model (CNN)	95%

6. CONCLUSION

The disease prediction system for Diabetes, Breast Cancer, Malaria, and Pneumonia represents a significant advancement in healthcare, utilizing machine learning and deep learning models to enhance diagnostic accuracy and efficiency. Integrated into a web-based platform, it aims to provide real-time, reliable diagnoses, especially in resource-limited settings with limited access to specialized equipment and trained professionals. Diabetes and Breast Cancer prediction models use machine learning algorithms to process structured data, enabling early detection and improved disease management, while Malaria and Pneumonia models employ Convolutional Neural Networks (CNNs) to analyse medical images, showcasing deep learning's power in image-based diagnostics. The system's real-time capabilities allow healthcare providers to make faster, more informed decisions, improving patient outcomes and reducing diagnostic delays and misdiagnosis. Despite challenges such as the need for high-quality datasets and computational resources, the potential impact on global health is immense, particularly in underserved regions, where early detection can lead to better treatment outcomes and reduced mortality. By offering an affordable, efficient, and scalable solution, the disease prediction system has the potential to make a lasting contribution to healthcare delivery, revolutionizing the way diseases are diagnosed and managed, and making healthcare more accessible, accurate, and efficient worldwide.

REFERENCES

[1]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University Computer and Information Sciences 25, 127–136. doi: 10.1016/j.jksuci.2012.10.003.

[2]. C. Christiansen-Jucht, P. E. Parham, A. Saddler, J. C. Koella, and M.-G. Basa´nez, "Temperature during larval development ˜ and adult maintenance influences the survival of Anopheles gambiae s.s," Parasites & vectors, vol. 7, 2014.

[3]. Y. A. Afrane, A. K. Githeko, and G. Yan, "The ecology of Anopheles mosquitoes under climate change: case studies from the effects of deforestation in East African highlands," Annals of the New York Academy of Sciences, vol. 1249, no. 1, pp. 204–210, 2012.

[4]. 6. M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in Systems Engineering (DeSE), Liverpool, 2016, P. 35-39.

[5]. P. Pratik, Hemprasad Patil, X-ray Imaging Based Pneumonia Classification using Deep Learning and Adaptive Clip Limit based CLAHE Algorithm (2020).

[6]. Domes Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[7]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

[8]. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE.

[9]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[10]. H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Blore, 2013, P. 1-5.

[11]. VIC, 2017, P. 642-645. S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, 10.1109/CEM.2017.7991863.

[12]. A.Davis, D., V.Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008). Predicting Individual Disease Risk Based on Medical History.

[13]. Adam, S., & Parveen, A. (2012). Prediction System For Heart Disease Using Naive Bayes.

[14]. Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. Information Technology Journal.

[15]. Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based on Medical History.

[16]. JyotiSoni, Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction.

[17]. K.M. Al-Aidaroos, A. B. (n.d.). K.M. Al-Aidaroos, A. B. (n.d.). 2012. Medical Data sssClassification With Naive Bayes Approach.

[18]. Nisha Banu, MA; Gomathy, B. (2013). Disease Predicting System Using Data Mining Techniques.