Machine learning-based automated medical diagnosis for healthcare

Harsh Khatter[1], Ankit Yadav[2], Ayush Srivastava[3]

[1,2,3]*Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, Uttar Pradesh, India*

[1]harsh.khatter@kiet.edu

[2]ankit.1923co1041@kiet.edu

[3]ayush.1923co1040@kiet.edu

Abstract— **A sizable segment of the world's population lacks access to quality healthcare. The success of healthcare ultimately depends on the doctor's skill. In this study, we investigate if this knowledge may be represented as an information corpus, or as data that has been retrieved using data mining methods, particularly the Machine Learning & Deep Learning Model, to make a diagnosis. When the medical diagnosis is made widely available, coverage increases, and life quality improves. In order to determine whether inferences about the causes of various diseases can be made from the data, this paper provides an overview of machine learning approaches used in the classification of various diseases. We outline a few of our findings from the trials we ran before offering some suggestions for the future.**

The difference between the current state of health and an acceptable or desirable health condition is the health problem. By lowering doctor visits, hospital stays, and diagnostic testing procedures, monitoring systems are designed to lower health care expenditures. Using the data mining modeling technique, the integration of clinical decision support with computer-based patient records could decrease medical errors, increase patient safety, stop unwelcome practice variance, and improve practice outcomes. Connecting patients and doctors through a user-friendly interface will make it easier for patients to use in emergency situations.

Keywords— Artificial Intelligence, Machine Learning, Health Analysis, Preventive systems

I. INTRODUCTION

Machine learning (ML) is a branch of artificial intelligence that uses "learning ability provided to computers without additional programming" to address problems in the real world. Research into whether computers could learn to emulate the human brain led to the development of machine learning. When Arthur Samuel created the first checkers game-playing software in 1952, ML made its initial attempts to generate the necessary skills to defeat a world champion. Later in 1957, Frank Rosenblatt developed an electronic system that can mimic the way the human brain works to learn how to handle complex issues [5].

The growth of ML has facilitated the increased use of computers in medicine.There are three primary goals of data mining that tend to be prediction, description, and presentation. Prediction involves some variables or fields in the dataset to predict unknown or future values of other variables of interest. Description focuses on finding patterns describing the data that can be interpreted by humans, on the other hand Presentation plays an important role to be easily understandable to humans.

Health issues are more prevalent today than they were 25 to 30 years ago due to urbanization and industrialization. Unexpected changes in the environment have a direct impact on health. As a result, more health issues emerge daily, necessitating daily health condition checks. This project, which is based on real-time implementation and is more informative and realistic, can be very useful in helping to diagnose medical conditions and comprehend the many model feasibility options.

The availability of medical facilities for people is limited under these unheard-of economic conditions, and it is even more so for those with specific requirements. People with other illnesses didn't even get an opportunity to access medical aid because the healthcare industry became so preoccupied with Covid-19 instances [20]. We are here to provide a model that will help alleviate the lack of facilities and poor medical management.

The goal of creating automated systems is to decrease the amount of time and money spent on health care by lowering the number of doctor visits, hospital stays, and diagnostic testing procedures. The second main method is supervised learning, which uses machine learning techniques to learn a function from a collection of training data. Prediction and generalization are two key requirements for supervised learning algorithm performance. The trained function ought to be

capable of accurately predicting the results for data that are not included in the training set. It should also serve as a model that generalizes to new data points and captures the underlying traits of the training data.

Machine learning is frequently used to categorize diseases, and scientists are becoming increasingly more interested in creating such systems for many disease diagnosis and tracking as well. Diabetes, cardiovascular disease & many more are the top 10 killers worldwide, according to the World Health Organization (WHO). According to a study from January 2017, CDs are the leading cause of death worldwide. In the list of the top 10 causes of death over the past 15 years, the world's worst disease, which claimed 15 million lives in 2015, holds the top spot.

In this paper, the authors focuses on the various methods studied on preventive health and how to analysis the patient's health.

II. Literature Survey

Biosignals from patients have been used in the past to construct AI systems for clinical decision assistance. Such organized clinical data includes unprocessed signals. a lack of sufficient background for appropriate interpretation, whereas clinical publications with unstructured free text contain comprehensive explanations of broader clinical situations.

Tamilselvan. P [1]. Monitors based on blood pressure and ECG readings are available. Reactions are kept the signals that these sensors transmit to the Using a signal conditioner and amplifier, Raspberry Pi because the signals' levels are low (gain), an amplifier is necessary (scu). The signals that these sensors transmit to the Raspberry Pi are processed by a signal conditioner and amplifier, which is required because the signal levels are low (gain). Using a circuit, the signal is amplified and sent to a Raspberry Pi. Linux is used to run the Raspberry Pi computer. The system operates like a minicomputer processing system. Here, patients' ECG and blood pressure are measured with the appropriate sensors. Moreover, it is monitorable on a computer's monitor. Utilizing a Raspberry Pi and monitoring from anywhere, Internet sources are used everywhere.

Vivek Datla, Sadid A. Hasan [2]. outlines our Knowledge Graph (KG)-system for based clinical diagnostic inference. We performed substantial testing on the MIMIC-III benchmark dataset, analyzing different parts of a clinical note. Results proved that the details of the current illness's history were relevant. The parts on prior health histories often offer the greatest insight. inference of a clinical diagnosis in comparison to all portions. Furthermore, we demonstrated that the KG-based system can perform admirably with a loose accuracy metric in comparison to the cutting-edge CMemNN model.

Hiroshi Sugimura, Kazuki Utsumi [3]. They suggest a method that superimposes online service information onto the daily environment. Natural behaviors like speech and gestures are used to operate the system. The system is put together using three input/output devices: a microphone for sound recognition, a camera for gesture detection, and a projector for information presentation. We provided a detailed account of the conception, design, execution, and assessment of a prototype system. The proposal system's value was then confirmed.

Fabio Santos, Filipe Silva and Petia Georgieva [4]. focused on the skin lesion diagnosis techniques integrated inside eHealth apps that help individuals and medical professionals and are clearly needed as the prevalence of skin cancer grows. Meanwhile, recent developments in deep learning techniques enable performance that is close to that of a dermatologist and has a large room for growth, outperforming previous approaches. Before putting such tools into use in the real world, issues like the need for large datasets or the high computing needs must be resolved because they negatively affect how well models function. However, effective methods like these reduce their impacts through transfer learning and data augmentation, according to research. Finally, it is anticipated that when more information on skin lesions is made publicly available, these difficulties will lose some of their significance.

Berina Ali [5] provides an overview of machine learning methods for categorizing CVD and diabetes using artificial neural networks (ANNs) and Bayesian networks (BNs). A comparative study was carried out on a few publications released throughout the time period, between 2008 and 2017. In a few

articles, multilayer feedforward neural networks using the Levenberg-Marquardt learning method are the most commonly utilized ANN type. Additionally, utilizing ANN improved the computation of the mean accuracy of observed networks, indicating a greater likelihood of obtaining more accurate findings for the categorization of CVD and/or diabetes.

Authors focus to check if seeding strategies have strong influences on the success of viral marketing campaigns or not. We can determine the era of marketing how one can have successful approach in getting viral videos [11].

III. METHODOLOGY

The methodology's goal is to forecast an individual's risk of developing kidney, lung, and breast cancer, heart disease, and diabetes using a few questions and machine learning models in an end-to-end procedure. the system I used for my research has the following software and system configurations: On an Intel(R) Core(TM) i5-2310M GPU @1650Ti with 8 GB RAM, Jupyter Notebook 5.5.0 and VS Code 1.73 are used to implement Python 3 and the Flask framework.

Figure 1 displays a block schematic of the fundamental procedures used for each machine-learning model. To transform the raw data into a form that can be used, data cleaning is done first. Data analysis is carried out after data cleansing to ascertain the significance of characteristics. Once a goal has been established, it is time to start gathering the data required for analysis. Your data team will be responsible with cleaning and sifting through the data once it has been gathered from all the required sources. Because not all data is good data, data cleansing is crucial during the data analysis process. Data mining, which is referred to as "knowledge discovery inside databases," is one method. In order to predict what will probably happen next in the future, predictive analysis looks ahead to the future. These methods are a component of inferential statistics, which is the act of examining statistical data in order to make inferences about the connections between various sets of data. We have to utilize the Train/Test methodology to evaluate the model's performance. The train/test approach is a way to gauge how accurate your model is. Because you divide the data set into two sets, a training set and a testing set, the method is known as Train/Test. 20% for testing, 80% for training. Using the training set, you train the model. Utilizing the testing set, you test the model. Create the model by training it. To test a model is to determine its correctness. After confirming that our model is sound, we can now begin making fresh value predictions.

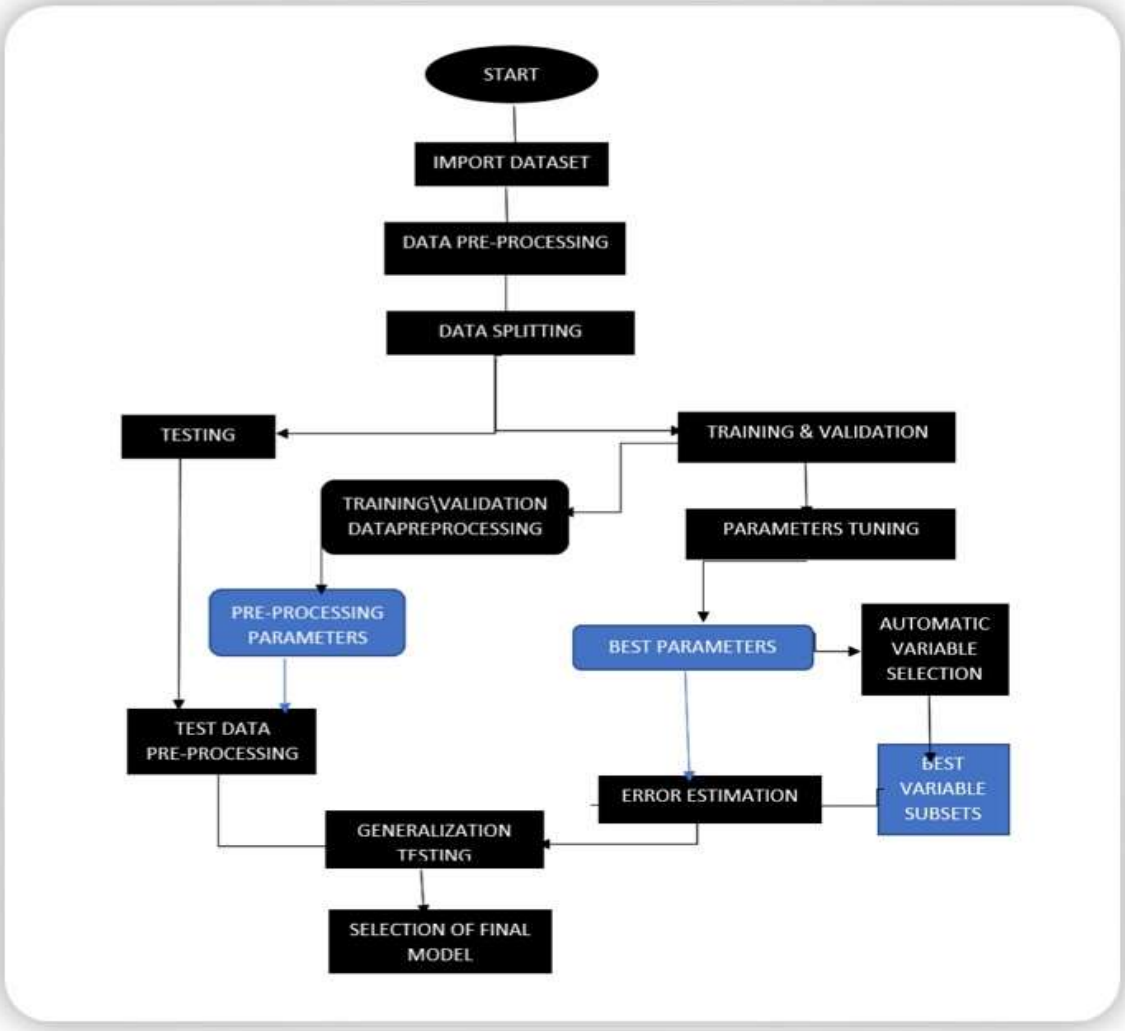

Figure 1: Flow chart of the proposed model

IV. ALGORITHMIC APPROACHES FOR PROPOSED WORK

A. *Random Forest Classifier Algorithm:*

A group of several decision trees is called a random forest. Decision trees are employed as parallel estimators in the bagging technique, which is used to construct random forests. When used in a classification issue, the outcome is determined by the majority vote of the findings from each decision tree. In a regression, the mean value of something like the target values in a leaf node serves as the prediction. The mean value of the decision tree outcomes is taken into account by random forest regression. [6,7]

B. *Logistic Regression:*

A set of independent variables is used to estimate discrete values (often binary values like 0/1) using logistic regression. By fitting data to a logit function, it aids in the prediction of an event's likelihood. Logistic regression is another name for it. The following techniques are frequently used to enhance logistic regression models: incorporate interaction terms, remove features, regularize approaches, and employ a non-linear mode. [12,13]

C. *Decision Tree:*

The Decision Tree method, a supervised learning technique used for issue classification, is one of the most widely used algorithms in machine learning today. Both continuous and categorical dependent variables may be classified using it. Based on the most important characteristics/independent variables, this method splits the population into two or more homogenous groupings.

Step 1: Gathering and identifying the data that will be provided to the network as input is the first stage in the categorization of diabetes or heart disease using Logistic Regression, Decision Tree & Random Forest. [14]

Step 2: The network receives a defined training dataset and the selected training method. After the training phase, the Logistic Regression, Decision Tree & Random Forest are also put to the test to get feedback on how well they categorize the condition.[15]

Step 3: The "training data set" is used to train the classifier, the "validation set" is used to fine-tune the parameters, and the "test data set" is used to evaluate the performance of the classifier. It's vital to keep in mind that only the training and/or validation set is available when the classifier is being trained. The test data set must not be used for classifier training. Only when the classifier is being tested will the test set be accessible.

Step 4: We may then create a confusion matrix, which indicates how effectively our model has been trained, after this is complete. True Positives, True Negatives, False Positives, and False Negatives are the four parameters that make up a confusion matrix. [16] In order to create a model that is more accurate, we would want to obtain more data for the true negatives and true positives. The number of classes has a direct impact on the size of the confusion matrix.

Step 5: The model creation process includes a step called model evaluation. Finding the model that best depicts our data and predicts how well the model will perform in the future is helpful. In order to enhance the model, we may adjust its hyper-parameters to attempt to raise its accuracy while also looking at the confusion matrix to try to increase the proportion of true positives and true negatives.

The dashboard will run the function that connects patients and doctors. The dashboard has been designed to get a better understanding of a patient's condition. The medical profession is being transformed by technological advancements, which is causing a rapid change in the medical business. Artificial intelligence (AI) is being used to diagnose illnesses, predict diseases, and develop treatments. In this article, it will be discussed how AI can support the management of level-three 911 calls and assist physicians in making better patient decisions. According to the doctors' available time slots, those who have appointments can show up for their appointments. To avoid any unneeded waiting times, this is done. Modern medical professionals are employing cutting-edge technology to provide more precise diagnoses and individualized treatment strategies [17,18]. Doctors can now give patients extensive medical data, an in-depth report on their symptoms, and personalized care plans that can be seen on a mobile device thanks to artificial intelligence (AI). With the use of an app, the doctor may view the patient's information, write him a prescription, and suggest testing. This software assists in quickly diagnosing the patient since it has all the information required to understand the ailment, its symptoms, and recommended treatments. The outdated paper lab slip is being replaced with a modern dashboard. Patients will be able to schedule an appointment at a time that works for them because these slips are now digital. As soon as the reports are finished being processed by the laboratory, they will also be automatically uploaded into the patient's database. Doctors use information from a variety of sources to provide the most precise diagnosis possible. With the more recent use of electronic medical records, a more complete image of patients is now possible, improving care. [19]

V. Discussion

Data mining techniques play a significant role in medical systems, which will significantly contribute to the advancement of the medical industry. This paper provides a disease categorization based on several data mining and artificial intelligence technologies. Furthermore, we discovered in the literature that there are two kinds of factors utilized in disease categorization.

We noticed in the literature that the classification of illnesses still has room for improvement. Despite the fact that proper characteristics may be derived from the ECG, it is a noninvasive strategy used to diagnose patients, and the ECG signal does not provide the necessary information. Because biosignals have an irregular structure, creating an effective technique for hidden factor extraction from

ECG signals is particularly challenging. Several studies have found that the feature extraction method is incapable of determining the exact values of unmasked ECG signal parameters. Furthermore, using a restricted dataset for classification may result in misclassification; hence, in order to overcome the error rate, it is critical to avoid using a short dataset for classification.

VI. CONCLUSION

In this project, the focus is on healthcare services, which are such an important aspect of our society, automating them relieves human stress while also making measurement simpler. Furthermore, the system's transparency promotes patient trust. When the threshold value is achieved, an alarm system consisting of a buzzer and an LED alerts the physicians, allowing them to respond more quickly.

This work focused on a variety of health concerns that affect not only India but the entire world. Visualize the dataset and identify diseases using all ML and DL algorithms. Integrate physicians and patients with an interactive and user-friendly interface so that patients may use it in an emergency. This model's three key purposes are prediction, description, and presentation. The patient's biometric data, which is captured, published online, and transferred to cellular devices, may be made available to scientists and researchers in medical disciplines in order to evaluate its value and reveal patterns, as well as for other research objectives.

Future research will focus on monitoring additional health-related indicators with a bigger collection of transducers, sensors, and correlation algorithms, as well as improving system dependability and resilience in the face of patient movement and connectivity losses.

REFERENCES

[1] D. Fagella, Applications of Machine Learning in Pharma and Medicine, 2017, [online] accessed on 14 Nov 2022.

[2] V. S. Pendyala and S. Figueira, "Automated Medical Diagnosis from Clinical Data," 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), 2017, pp. 185-190, doi: 10.1109/BigDataService.2017.14.

[3] F. Santos, F. Silva and P. Georgieva, "Automated Diagnosis of Skin Lesions," 2020 IEEE 10th International Conference on Intelligent Systems (IS), 2020, pp. 545-550, doi: 10.1109/IS48319.2020.9200090.

[4] V. Datla et al., "Automated clinical diagnosis: The role of content in various sections of a clinical document," 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 1004-1011, doi: 10.1109/BIBM.2017.8217794.

[5] J. Siddiqi, B. Akhgar, A. Gruzdz, G. Zaefarian and A. Ihnatowicz, "Automated Diagnosis System to Support Colon Cancer Treatment: MATCH," Fifth International Conference on Information Technology: New Generations (itng 2008), 2008, pp. 201-205, doi: 10.1109/ITNG.2008.62.

[6] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), 2017, pp. 1-6, doi: 10.1109/ICTKE.2017.8259629.

[7] R. Chellappa, S. Theodoridis and A. van Schaik, "Advances in Machine Learning and Deep Neural Networks," in Proceedings of the IEEE, vol. 109, no. 5, pp. 607-611, May 2021, doi: 10.1109/JPROC.2021.3072172.

[8] F. Ertam and G. Aydın, "Data classification with deep learning using Tensorflow," 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 755-758, doi: 10.1109/UBMK.2017.8093521.

[9] H. Božiković and M. Štula, "Web design — Past, present and future," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 1476-1481, doi: 10.23919/MIPRO.2018.8400266.

[10] D. S. Blyth, "Web page design for HTML (and friends)," IPCC 96: Communication on the Fast Track. IPCC 96 Proceedings, 1996, pp. 89-103, doi: 10.1109/IPCC.1996.552585.

[11] H. Ran, W. Zhuo and X. Jianfeng, "Web Quality of Agile Web Development," 2009 IITA International Conference on Services Science, Management and Engineering, 2009, pp. 426-429, doi: 10.1109/SSME.2009.112.

[12] S. Drobi, "Play2: A New Era of Web Application Development," in IEEE Internet Computing, vol. 16, no. 4, pp. 89-94, July-Aug. 2012, doi: 10.1109/MIC.2012.84.

[13] H. Sugimura et al., "Development of immersive display system of web service in living space," 2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE), 2014, pp. 49-50, doi: 10.1109/GCCE.2014.7031106.

[14] N. Sandhya and K.R. Charanjeet, "A review on Machine Learning Techniques", International Journal on Recent and Innovation Trends in Computing and Communication, pp. 395-399, 2016, ISSN 2321-8169.

[15] Harsh Khatter, Amit Kumar Gupta, Ruchi Rani Garg, and Mangal Sain. 2022. "Analysis of the S-ANFIS Algorithm for the Detection of Blood Infections Using Hybrid Computing" Electronics 11, no. 22: 3733.

[16] Kanika Gupta, Nandita Goyal, Harsh Khatter "Optimal reduction of noise in image processing using collaborative inpainting filtering with Pillar K-Mean clustering", The Imaging Science Journal (2019), Vol. 67(2), pp:100-114, ISSN 1743-131X

[17] Sharma, Sarang; Gupta, Sheifali; Gupta, Deepali; Juneja, Abhinav; Khatter, Harsh; Malik, Sapna; Bitsue, Zelalem Kiros "Deep Learning Model for Automatic Classification and Prediction of Brain Tumor", Journal of Sensors (2022), Hindawi, Vol 2022(3065656), 8 April 2022.

[18] Gunjan Raghav, Harsh Khatter, "Intelligent Curation Fuzzy Inference System for Blood Infections in Android", IEEE 4th International Conference on Computational Intelligence & Communication Technology (CICT), 2018, Ghaziabad, India.

[19] Vaishali Aggarwal, Harsh Khatter, Anil Kumar Ahlawat, "Performance analysis of the competitive learning algorithms on Gaussian data in automatic cluster selection", 2nd IEEE International Conference CICT 2016, Feb 12-13 2016, ABES EC Ghaziabad.

[20] Anjali Jain, Harsh Khatter, Rajesh Kumar Tewari, (2020) "Insight To Pandemic Covid-19: Analysis, Statistics and Prevention", Aegaeum Journal, Vol. 8 (7), pp: 551-561, ISSN 0776-3808 DOI:16.10089.AJ.2020.V8I7.285311.3758

[21] Harsh Khatter, Amit Kumar Gupta, Ruchi Rani Garg, and Mangal Sain. 2022. "Analysis of the S-ANFIS Algorithm for the Detection of Blood Infections Using Hybrid Computing" Electronics 11, no. 22: 3733, 14 Nov 2022, https://doi.org/10.3390/electronics11223733

[22] Harsh Khatter, Anil K Ahlawat, "An intelligent personalized web blog searching technique using fuzzy-based feedback recurrent neural network". Soft Computing, 24, 9321–9333, 29 April 2020, (2020). https://doi.org/10.1007/s00500-020-04891-y

[23] Harsh Khatter, Anjali Jain, Poonam Pandey, "Classification and Categorization of Blood Infection using Fuzzy Inference System", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-5 Issue-2, May 2015, pp. 95-97