Non-communicable Disease Detection Based on Early Symptoms Using Machine Learning Approach Enabling Smart Healthcare Model (IoMT)

Tasnim Binte Shiraj[1], Ali Mortuza[2], Kazi Md Anisur Rahman[3], Tajbia Karim[4]

Abstract

Early detection of disease can prevent fatality and even save the lives of individuals. Since many diseases may have some common symptoms, it is essential to critically analyze symptoms for the correct prediction of diseases. Machine learning influences disease prediction, analyzing numerous features with high accuracy. In our country, elderly people suffer mostly alone as every other member remains busy outside the home, so they lack proper care and constant observation. A cloud infrastructure that allows digital devices to gather, infer, and exchange health data is called the Internet of Medical Things (IoMT). As the global economy grows, so will the cost of linked healthcare. The ever-lowering cost of sensor-based technologies is the reason behind the extraordinary expansion of IoMT. This paper reviews which machine learning algorithm is most suitable for detecting non-communicable diseases in terms of precision, specificity, accuracy, and confusion matrix. It is possible to keep track of old persons by detecting disease from the early stages of symptoms. We used OHAS (Occupational Health Automated System) dataset for finding the accuracy of the disease detection system. We utilized several machine learning techniques for detecting non-communicable diseases (for example, K-Nearest Neighbor, Decision Tree, Support Vector Machine (SVM), XG-Boost, and logistic regression). This article's objective is to investigate the repercussions of using the aforementioned algorithms effectively and find out which is the best algorithm for early Disease detection. We observed that from the mentioned algorithms, XG-Boost outperforms all other algorithms and gives the best accuracy of 86.24 percent.

Keywords: Dataset, Classifier, Preprocessing, Feature Selection, XGBoost, SVM.

1. Introduction

A component of the Internet of Things is called the Internet of Medical Things (IoMT) which deals with the gathering, processing, data exchange, and storage for

[1,2,3,4]Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh

Email: tasnimarrhiza512@gmail.com[1],alibinabdulwadud@gmail.com[2],anisurtopu555@gmail.com[3], tajbia.karim@bup.edu.bd[4]

medical purposes via a network of specific devices (things) designed to assure patient safety and data security. The IoMT platform is based on a common IoT platform. During the pandemic period, we face a major issue regarding medical treatment. It was really tough for us to visit the hospital and clinic on a regular basis. Also, various restrictions were a concern. To overcome this problem, we can build potable IoMT devices that can easily and accurately detect our diseases. On the other hand, many of us are not willing to go to the hospital for every single health issue. This type of negligence becomes the reason for critical health issues and diseases. In this circumstance, IoMT can play a big role. Like as we have a device that is able to detect diseases at a very early stage that might save one's life. It will help the patient to take proper decisions and clinical steps.

In recent times, various design options have been demonstrated using both existing protocols and regularly utilized bands. Various algorithms were utilized, including SVM, Logistic Regression, J48, KNN, Decision Tree, and Nave Bayes. Regression and classification are the two main applications of Support Vector Machine or SVM. Even when the data is not otherwise linearly separable, SVM categorizes data points by mapping them to a high-dimensional feature space. The data is processed such that the division may be shown as a hyperplane once the categories have been divided. Logistic regression is a technique for estimating the probability of a discrete output given an input variable. The most typical logistic regression model has a binary result.

In heart disease detection, using logistic regression, they get 87.1% accuracy. The supervised machine learning algorithm KNN (K-Nearest Neighbor) is used for classification and regression problems. The symbol "K" stands for the quantity of new unknown variables that must be predicted or classified and whose nearest neighbors must be counted.' As a result, the KNN algorithm can be employed in high-precision applications. The accuracy of the predictions is influenced by the distance measure. The KNN approach is best for applications that require a lot of domain expertise. J48 is a prominent machine learning method that can categorize and continually identify data. It consumes more memory and decreases the performance and accuracy of medical data classification when used, for example. J48 had the highest accuracy of 99.52 percent for identifying dementia in this literature review research. Although it may be used to solve classification and regression issues, the decision tree is most frequently employed to tackle classification difficulties. A probabilistic classifier with strong independence assumptions, the Naive Bayes method is based on probability models. This entire

technique is now very popular and is being used in a variety of studies. Doctors and medical institutions are also affected and are interested in these algorithms based IoMT devices for better treatment.

The main objectives of the paper are:

- To investigate the background and previous work in this field and find out which approach they took up and what they found in their research findings.
- To characterize the illnesses using various calculations such as XGBoost, K-Nearest Neighbor, Decision Tree, and Support Vector Machine.
- To find the most dangerous threat variables that cause these disorders.
- Comparison of various arrangement processes and identification of the most appropriate characterization strategy for provided data.

2. Literature Survey

Shokat Ali et al. (2020) presented Roles, Challenges, and Applications, which are based on the usage of IoMT guidelines and instruments has radically revolutionized orthopedic medical treatment, procedures, and services during the COVID-19 pandemic. Data gathering, report monitoring, patient database access, analysis of test pictures, and other tasks or features are all made possible by the IoMT system. Another issue to consider is interoperability. Rita Chhikara1 et al. (2018), proposed Analysis of Different Machine Learning Algorithms in Comparison for Dementia Detection. Dementia affects 46.8 million people now, with a projected 74.7 million more in 2030 and 131.5 million more in 2050. SVM, Logistic Regression, J48, and Nave Bayes were employed in this case. J48 has the highest accuracy of 99.52 percent. Good accuracy, specificity, discriminant power, and sensitivity are characteristics of the linear discriminant analysis said to Professor Pagar et al. (June 2021). They developed a model using the NB (Naive Bayesian), Decision tree map, Random Forest of Machine learning approach with the AES (Advanced Encryption Standard) algorithm, and SHDP (Smart Heart Disease Prediction) to address the issue of heart disease prediction. Heart disease may be identified with 87.1% accuracy using Logistic Regression. Diabetes prediction using Support Vector Machine (a linear kernel) was 85.71%, while Breast Cancer detection using AdaBoost classifier was 98.57 percent. Joyia et al. (2017) put forth The Applications, Benefits, and Future Challenges of Internet of Medical Things (IOMT) in the Healthcare Domain. The Internet of Things (IoT) is the most promising solution for the healthcare business, allowing individuals to control their own ailments and obtain aid in the event of an emergency. The Smart Rehabilitation System, Kidney Abnormality Detection System based on IoT Using Ultrasound Imaging, the Patient Physiological Monitoring System, Patient Posture

Recognition Application Using Supervised Learning, Safe and Intelligent Medical Healthcare System. Mohammad et al. (2020) suggested a healthcare monitoring

system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. This is based on the IoMT cloud environment's heart disease diagnostics. This employs using a neuro-fuzzy adaptive inference system (ANFIS) and MSSO (Modified Salp Swarm Optimization). MSSO-ANFIS enhances search capabilities by utilizing the Levy flight algorithm. ANFIS is a gradient-based learning algorithm that is prone to get stuck in local minima. MSSO-ANFIS has a greater accuracy of 99.45 percent with 96.54 precision than the other approaches. People may forecast heart disease by preprocessing (replacing missing values, data normalization) and feature selection. Taisa D et al. (2019) made a proposal called Breathing Monitoring and Pattern Recognition with the Worn Sensor, which is based on breath monitoring and pattern recognition using a wearable sensor. The nose, airways, lungs, etc. made the respiratory system. Many respiratory wearable sensors are used in this study, comprising acceleration sensors, oximetry sensors, humidity sensors, pressure sensors, and acoustic sensors. In this work, certain signal processing methods (amplification, filtering, analog to digital conversion, Fast Fourier transform (FFT)) are employed. This study employs Naive Bayes, SVM, and Artificial Neural Networks algorithms.

Danial H et al. (June 2017) proposed Diagnosis of Dementia from Early Clinical Data based on employing machine learning algorithms to identify dementia from clinical data. Based on the MMSE-KC data, data preprocessing will be performed. This study employs the Support Vector Machine (SVM), Naive Bayes, Logistic Regression, and bagging method. (SVM) is a data analysis and pattern recognition mapping model with an accuracy of roughly 87 percent. The FCM and PNN accuracy in this model were 74% and 69%, respectively. The Naive Bayes model was about 70% accurate. Bagging also aids in the reduction of variance and the prevention of overfitting. Using TriAxial Accelerometer Sensors, a Dementia Wandering Detection and Activity Recognition Algorithm was suggested by Kyu-Jin et al. (2009). This uses a triaxial method to identify dementia wandering where they offer an effective approach for translating raw sensor data into readable patterns in order to classify current activities, and then comparing these patterns to previously recorded patterns to detect many problematic patterns, such as wandering, one of the early indicators of dementia, and so on. The sensor node is made up of MSP430 for MCU, CC2420 for wireless connection, six 3-axis acceleration sensors, and three MUXs to select and receive data from sensors. Haruka et al. proposed An Early Detection System for Dementia Using the M2M/IoT Platform (2016). This system recognizes the signs of old people living

alone by using the M2M (Machine-to-Machine) / IoT (Internet of Things) platform. The authors devised three different sorts of analytical procedures. One way was behavioral data analysis/group comparison. Secondly, the way was behavioral data analysis/comparison with past behaviors. Another way was behavioral data and attribute data analysis utilizing Discriminant Analysis Method and Multiple Logistic Regression.

Radhika et al. (2020) discussed the use of decision trees and electronic health record analysis to predict diseases based on symptoms where they attempt to forecast ailments of users based on their symptoms. They employed the Decision Tree Classifier to attain their goal, which aids in detecting the patient's health state after obtaining their symptoms by providing the anticipated illness. To offer a summary of the health record, they employed NLTK - Natural Language Toolkit libraries. Their suggested system comprises two modules: one for illness prediction and the other for health records. Muhammad et al. (2021) observed the efficiency with which the smart healthcare (SHC) model can be used to monitor older persons testing aboard the IoMT dataset. Their proposed model used a way to monitor elderly persons using artificial neural networks (ANNs) which achieved 0.936 accuracy. Recently, there has been a surge of interest in incorporating machine learning into CPS, which can aid in disease categorization, detection, monitoring, and prediction for a variety of NCDs. As an example of NCDs, an original machine learning-based CPS for health is provided that successfully analyses data from wearable IoT sensors for diabetes risk prediction early on. Following tests with several machine learning methods, the researchers discovered that the Random Tree approach has the highest precision, needs the least amount of time to develop a model, and has a 94 percent accuracy in predicting the likelihood of diabetes at an early stage (Rahatara et al. (2021). A model for predicting multiple diseases based on Symptoms using Machine learning was proposed by the authors Talasila et al. (2021). The goal of this effort was to use Machine Learning (ML) models to help clinicians predict and analyze diseases at an early stage. They utilized a dataset that includes 4920 patient records that were judged to have 41 diseases. This research investigates three models' execution of a clinical record, yielding Decision Tree (0.973154), LightGBM (0.973154), and Random Forest (0.98315). Rinkal et al. authors created a technique for predicting diseases based on numerous machine learning techniques. There were more than 230 diseases in the dataset that was used for processing. They examined the dataset using Gaussian Nave Bayes, Fine, Medium, and Coarse KNN, Weighted KNN, Fine, Medium, and Coarse Decision Trees, Sub Space KNN, and RUSBoosted trees are among the ML models used. Among them, Weighted KNN gave the best result for the disease prediction model which had an accuracy of 93.5%. The authors Hamsagayathri et al. searched for the best ML techniques that are found to detect different specific

diseases. They observed various Computer Aided Diagnostic tests to examine ML techniques which results better. The authors found The Nave Bayes technique can be used to diagnose diabetes. SVM is useful for accurately diagnosing heart disease with an accuracy of 94.60%. For Liver Disease K Star resulted best with 83.47% accuracy and for Dengue disease DT, ANN, and RS resulted best with 99.96% correctness. The suggested deep learning system, Laavanya (2019) was trained where it uses 26,000 samples per dataset and obtains a 99.7% accuracy rate. As a result of this research, they built an accurate, rapid identification of stress levels As a result of this research, The "Stress-Lysis" IoMT system was developed, which may measure user-end stress levels (at the edge) and save the data in the cloud. To measure stress levels in real-time, the suggested Sensor device for stress and lysis may be readily fitted into a palm band or a glove. The proposed idea behind the IoMT-based stress detection system can provide monitoring of both acute and chronic stress. Deep Neural Networks (DNN) or Deep Learning Models are employed in huge dataset pattern recognition applications. The illness prediction system Sneha et al. (2020) constructed using a Decision Tree classifier, Random Forest classifier, and Naïve Bayes classifier among some other machine learning techniques is demonstrated in this study paper. This research gives a thorough comparison of three algorithms' performance on a medical record, each with an accuracy of up to 95%.

The suggestion with an improved linear model, recursion increased random forest (RFRF-ILM), Chunyan et al. (2020) to diagnose heart disease is described in this study. The purpose of this article is to elicit the main aspects of the prediction of cardiovascular disease using machine learning techniques. To cluster datasets, Decision Tree (DT) feature variables and criteria are utilized. The classifier is then used to estimate the performance of each data set. Because they have a lower error rate, the best performing models are chosen based on the findings. The suggested RFRF-ILM approach is used to combine the characteristics of the linear model with the random forest. The RFRF-ILM predicts cardiac disease with good accuracy. The suggested approach reduces diagnostic costs and time while increasing treatment accuracy. Dhiraj et al. (2019) provided a general illness prediction based on the patient's symptoms. They use the Machine learning techniques K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN) for accurate sickness prediction. The accuracy of CNN-based general sickness prediction is 84.5 percent, which is higher than the accuracy of the KNN approach. They begin by downloading a UCI's machine learning website which has a disease dataset. Following preprocessing, the feature was retrieved and chosen. The data is then classified using classification algorithms such as KNN and CNN. They can

accurately forecast illness using machine learning. A sample of 4920 patients' records with 41 disorders was chosen for study. 41 illnesses made up the dependent variable. 95 of 132 independent variables (symptoms) associated with illnesses were chosen and optimized. Laavanya et al. suggested a novel stress detection technique called iStress During Physical Exercise (2018), it measures stress levels by measuring body temperature, the pace of motion, and sweat. They constructed the Fuzzy type controller of the Mamdani type which gave accuracy as 97% accuracy in assessing a person's stress level. Peiying et al. (2019), proposed a system where the sentence structure is assessed by building a syntactic tree to extract the subject, predicate, and the object of the sentence from raw input sentences that have been evaluated by the syntactic analyzer to determine if the word sequence is lawful or not. Once the sentence has been preprocessed, it is sent into word2vec, which creates a vector representation of the sentence as input for CNN. Finally, they apply the Manhattan distance formula to calculate the sentence vector output's similarity score. They employed the SPO model to extract symptoms information as the neural network's input, then processed the model's output sentence vector using the Manhattan distance algorithm to generate the most similar disease prediction findings. The SPO model has a 75.8% accuracy rate, SPO outperforms previous models, demonstrating that it can better capture phrase meaning. Currently, medical applications employ body area networks, RFID, and Bluetooth. Blood pressure, ECG signals, and EMG activity are all monitored via sensors embedded in garments. With Optimized Neural Networks, Intelligent Guided Particle Local Search (IGPLONN) technique is suggested by Mohamed et al. (2020) based on IoMT. The suggested system's performance has been experimentally validated using MATLAB to ensure its superiority. When compared to existing approaches, the suggested IGPLONN based on the IoMT method achieves the highest accuracy of 98.3 percent. The system employs the backpropagation network throughout this procedure. The system's quality is measured using MATLAB results, which show that the suggested IGPLONN approach has the highest accuracy of 98.3% after comparing it to other approaches.

3. Framework of ML based Disease Detection

Figure 1: Data collection to result flow in ML framework

3.1 Data Collection

Data was gathered from the internet to diagnose the condition. Only the true symptoms of the condition were collected; no dummy values were used. The disease's symptoms were gathered from kaggle.com. The dataset that we used for

this research article was OHAS Dataset. This csv file has 2129 rows of patient records with their symptoms (a variety of symptoms) and the condition that corresponds to them (148 a disease classification).

3.2 Data Pre-processing

Pre-processing data is a data mining technique that entails modifying or re-encrypting raw data to create an analysis-ready structure. It is known that efficiently the knowledge methodologies for pre-processing employed in provided work were deciphered using calculation as pre-processing of information, which included the following.

3.3 Data Purification

Certain measures are used to purify data. Along these ideas, settling the debt is similar to stuffing in lost value information with inconsistencies.

3.4 Data Scaling and Removing Redundancy

The dataset had some empty fields in some columns. We gave some random values to them (In the weight and height column) and some rows did not have symptoms which we removed for making the dataset workable. We also scaled the column’s value.

The Kaggle repository provided the Disease/Symptoms database. It has numerous features (symptoms) and classes, as previously stated (diseases). To train the model, we use this to construct training and testing sets. We collect the user's symptoms and apply the trained algorithm to forecast the disease. A summary of the health report, on the other hand, is prepared based on the patient's record. Symptoms that are extremely important in relation to a specific condition. This is done in order to spread the sickness. In most cases, any prediction system will only look at a preset platform. However, with a low confidence level, this would only partially find the result. In order to arrive, we employ health record analysis in this paper to deliver individualized input, which offers us a higher level of confidence as well as system interaction with the user. We split the data of the dataset into two different sets. Training module set and testing the module set. We split this by a proportion of 80:20 which is a decent splitting done to design a prediction model. Disease prediction is one module of the total system. The other is linked to a medical record. The second lesson is intended to boost your confidence. To

achieve the best results, one of the training data sets was chosen. The user is involved in this process interface for gathering user input and informing them of their condition.

4. Methodology

The dataset we examined has a combination of stages that result in 148 diseases. This dataset contains 2129 documents from distinct patient samples, giving us a wide range of illness combinations of gender, height, weight, BMI Levels, Severity, and so on.

- The main goal of our approach is to enable elder citizens to easily access information about their health concerns without the assistance of technology or medical professional, even if they have no prior knowledge of the medical field.
- The fundamental idea behind how the algorithm generates more accurate projections is machine learning. The ANN or CNN or LSTM technique outperforms the medical field estimation in a comparable scenario.

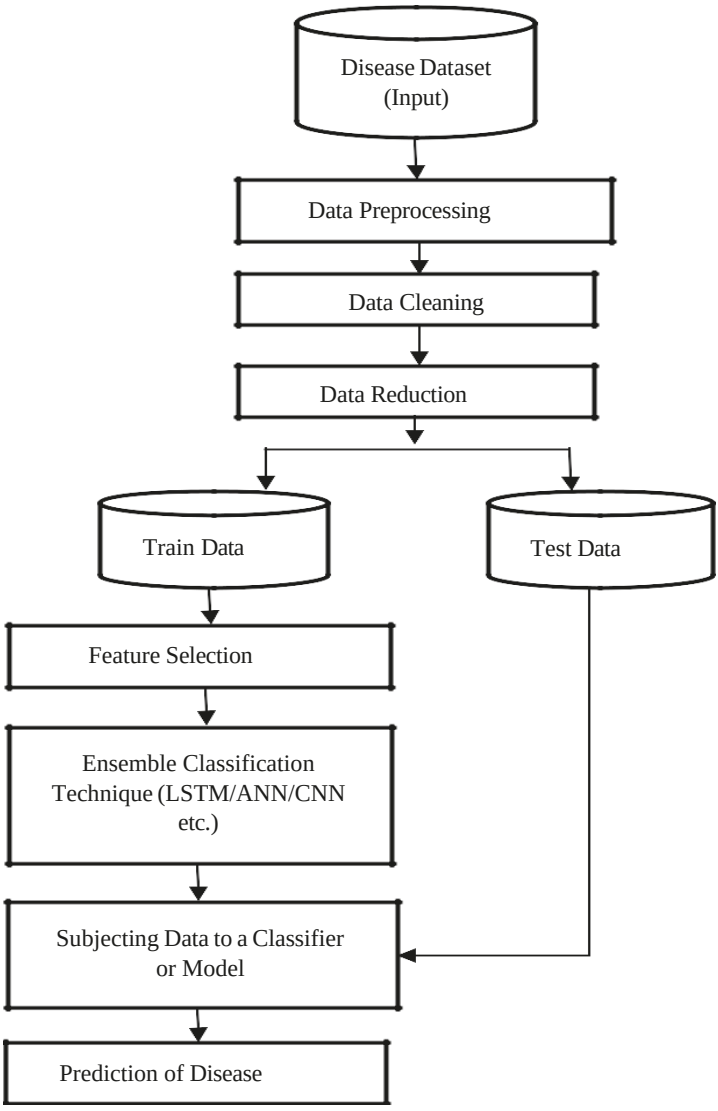

Figure 2: Methodology flowchart

A. Inputs (Patient Symptoms): When designing the algorithm, we assumed that the user would have a clear idea about the symptoms he is experiencing.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Disease	Disease_CUI	Symptoms	Symptom_CUI	Weight	Height	Intensity	Severity	Age	Gender	BMI_Leve	Region	Season
2	influenza	C0162565	uncoordi162ti	C0039239	68	180	high	medium	24	female	27.9	southwes	Summer
3	influenza	C0162565	fever	C0000737	68	170	low	medium	23	male	33.77	southeast	Summer
4	influenza	C0162565	pleuritic pain	C0235704	68	162	low	low	24	male	33	southeast	Summer
5	influenza	C0162565	snuffle	C0030554	68	162	high	medium	34	male	22.705	northwes	Summer
6	influenza	C0162565	throat sore	C0030552	68	185	low	high	21	male	28.88	northwes	Winter
7	influenza	C0162565	malaise	C0020538	68	185	medium	medium	21	female	25.74	southeast	Winter
8	influenza	C0162565	debilitation	C0020555	68	185	medium	medium	25	female	33.44	southeast	Winter
9	influenza	C0162565	asthenia	C0005758	68	185	low	high	25	female	27.74	northwes	Winter
10	influenza	C0162565	chill	C0030252	68	185	medium	low	27	male	29.83	northeast	Winter
11	influenza	C0162565	scleral icterus	C1868917	68	185	high	medium	27	female	25.84	northwes	Winter
12	influenza	C0162565	162sal flaring	C0392684	68	188	medium	high	31	male	26.22	northeast	Winter
13	influenza	C0162565	dysuria	C0012833	68	188	medium	low	31	female	26.29	southeast	Winter
14	influenza	C0162565	lip smacking	C0003467	68	188	high	high	31	male	34.4	southwes	Winter

Figure 3: Normal input data

B. Inputs (Preprocessed Dataset)

	A	B	C	D	E	F	G	H	I	J
1	Disease	Symptoms	Weight	Height	Intensity	Severity	Age	Gender	BMI_Leve	Season
2	influenza	uncoordi162tio	68	180	high	medium	24	female	27.9	Summer
3	influenza	fever	68	170	low	medium	23	male	33.77	Summer
4	influenza	pleuritic pain	68	162	low	low	24	male	33	Summer
5	influenza	snuffle	68	162	high	medium	34	male	22.705	Summer
6	influenza	throat sore	68	185	low	high	21	male	28.88	Winter
7	influenza	malaise	68	185	medium	medium	21	female	25.74	Winter
8	influenza	debilitation	68	185	medium	medium	25	female	33.44	Winter
9	influenza	asthenia	68	185	low	high	25	female	27.74	Winter
10	influenza	chill	68	185	medium	low	27	male	29.83	Winter
11	influenza	scleral icterus	68	185	high	medium	27	female	25.84	Winter
12	influenza	162sal flaring	68	188	medium	high	31	male	26.22	Winter
13	influenza	dysuria	68	188	medium	low	31	female	26.29	Winter
14	influenza	lip smacking	68	188	high	high	31	male	34.4	Winter

Figure 4: Preprocessed data

5. Implementation

Five classifiers are used to calculate information for the Disease Prediction Framework. These, for instance, were XG-Boost, Decision tree, K-Nearest Neighbor, Logistic Regression, and SVM. After that, we will try to create a model using ANN, CNN, or LSTM to classify diseases with a better accuracy rate than these classifiers.

5.1 Decision tree Model

The algorithm's order functioned as a Decision tree, which resembles a tree with multiple branches. As a result, it separates the dataset into smaller and smaller subgroups, resulting in the prediction of an objective value by examining the arrangement of unambiguous suppositions and then regulating on highlight esteem (manifestations for our circumstance) (disease). The most significant sections of a tree are the decision Node and the Leaf Node.

- Decision Node: A decision node is a node that has at least two branches. In our research, each manifestation is handled as a decision node.
- Leaf Node: The order is made up of leaf nodes, indicating that the decision can come from any branch. As a result, the node represents sickness on a tree.

5.2 XGBoost Model

An ensemble learning strategy is XGBoost. Relying just on the output of a single machine learning model might not always be sufficient. The prediction potential of several learners may be systematically combined through the use of ensemble learning. A single model that integrates the output of many other models is the ultimate result.

5.3 K-Nearest Neighbor

The k-nearest neighbors (KNN) technique is a data categorization method that uses the data points closest to it to estimate the likelihood that a data point belongs to one of two categories. To address classification and regression difficulties, the supervised machine learning technique k-nearest neighbor is applied. However, it is mostly employed to solve categorization issues.

5.4 Logistic Regression

The risk of categorization issues with two possible outcomes is modeled using logistic regression. It's an expansion of the classification issue with the linear regression model. When the dependent variable (target) is categorical, logistic regression is utilized.

As an illustration:

- To identify spam in an email (1) or (0)
- Whether or not the tumor is cancerous (1) or (0)

5.5 Support Vector Machine (SVM)

The Support Vector Machine, or SVM, is a well-liked Supervised Learning technique for dealing with classification and regression problems. The goal of the

SVM algorithm is to identify the ideal decision boundary or line for classifying n-dimensional space into groups so that the following data points may be quickly assigned to the appropriate category. A hyperplane is a mathematical term for the best decision boundary.

The extreme vectors and locations that will help build the hyperplane are chosen using SVM. Because support vectors are utilized as extreme instances, the method is known as an SVM classifier.

6. Results

After reviewing 24 papers we found several algorithms that used IoMT for early disease detection. For a hand on experience and examination, we chose the OHAS dataset and implemented Logistics Regression, Support Vector Machine, K-Nearest Neighbor, Decision tree, and XGBoost. We implemented these algorithms to understand the in-depth working procedure of mentioned algorithms. Nonetheless, when compared to the other models, XGBoost performs much better. Each model's accuracy score is listed below:

Table-1: Algorithms Vs Accuracy Score

Classifier/ Algorithm	Accuracy Score
XG-Boost	86.24%
KNN	70.276%
Decision Tree	54.95%,
Logistic Regression	58.164%
SVM (Linear Kernel)	68.92%
SVM (Polynomial)	65.58%

After preprocessing and data cleaning we imported the necessary libraries and implemented our dataset in different classifiers. We had a dataset containing 2129 rows from distinct patient samples, providing us with a huge variety of disease combinations based on gender, height, weight, BMI level, severity, and other factors. We may deduce from these results that each of the five Models performs excellently on the dataset.

7. Discussion

The paper's major goal was to look into the history and prior work on this subject to see which strategy they chose and what they discovered in their study findings. To describe the diseases using XG-Boost, Computations of K-Nearest Neighbor, Decision Tree, and Support Vector Machine. To identify the most serious risk factors that contribute to these disorders. Multiple arrangement methods are compared, and the best characterization methodology for the data supplied is chosen.

Data was gathered from the internet in order to diagnose the disease. There were no dummy values included, only the genuine symptoms of the disease were recorded. The signs of the disease were found on kaggle.com. The OHAS Dataset was the source of the data for this study. This csv file has 2129 rows of patient data, each of which includes their symptoms (a wide range of symptoms) and the condition that correlates to them (148 a disease classification). For processing the data, we followed three steps mainly, they are- Data Cleaning, Data Scaling, and Redundancy Removal.

The Kaggle source provided the Disease/Symptoms database. It has a range of traits (symptoms) and categories, as previously stated (diseases). This is used to create the model's training and testing sets. The symptoms of the user are collected, and the trained model is utilized to predict the ailment. On the other hand, depending on the patient record, a summary of the health report is prepared. Symptoms that are critical concerning a certain disease. This is done for the disease to spread. The dataset we worked on consists of a mix of phases that result in 148 diseases. In consideration of the 2129 documents of distinct patient samples in this dataset, we have a vast diversity of disease combinations with gender, height, weight, BMI Levels, Severity, and other variables.

For information calculation, the disease prediction framework uses five classifiers. These included XGBoost, Decision Tree, K-Nearest Neighbor, Logistic Regression, and SVM, to reference a few. We loaded the appropriate libraries and used our dataset in several classifiers after preprocessing and cleaning the data. This csv file dataset contains Symptoms, Diseases, Intensity, Severity, Gender, Season, etc. in Categorical form means the columns are in string representation, which is the Object datatype in the Pandas Data frame. Because categorical features are string data, humans can easily interpret them. Machines, on the other hand, are unable to assess categorical data rapidly. As a result, categorical data must be translated into machine-readable numerical data. Machine learning models are unable to interpret categorical data. As a result, the conversion to numerical representation is required. There are numerous methods for converting category

data to numerical data. To convert categorical data, we'll use the Label Encoding method.

We found the result for each implemented individual classifier which are, in XGBoost accuracy level was 86.24%, for KNN it was 70.276%, in the Decision tree algorithm it was, 54.95%, in Logistic Regression algorithm it was 58.164% and for the SVM linear kernel, the training accuracy was 68.92%, whereas test data achieved 64.86% accuracy. In the SVM Polynomial kernel, the training accuracy was 65.58% whereas we got 62.29% for test data accuracy.

8. Future Research Directions

Now we are trying to create a model using ANN, CNN, or LSTM to classify disease with a better accuracy rate than XGBoost, KNN, and Logistic Regression classifiers.

8.1 ANN

The learning methods used by artificial neural networks (ANNs) have the potential to adapt or learn on their own when new data is received. They, therefore, provide a great modeling tool for non-linear statistical data. ANN epochs (iterations) are made up of Forward and Backward Propagation.

The input layer receives information and passes it on to the hidden levels. Each input neuron is given weights and bias through the link between these two layers, and the weighted total, which combines the two, is then sent through the activation function. Each input neuron is initially given a random number of weights. The result is calculated by the Activation Function after it chooses which node to fire for feature extraction. This is known as forward propagation, Weights are updated after comparing with the original output, and error is known to minimize the error rate which is known as Backward Propagation.

8.2 LSTM

Recurrent neural networks known as Long Short-Term Memory (LSTM) networks are capable of learning order dependency in problems involving sequence prediction. This tendency is crucial in complicated problem areas like machine translation, pattern recognition, and other fields. This can be used in our model to recognize the same symptoms as a pattern and detect the correct disease every time and diagnose patients with what can be done for that specific disease. As we are willing to make an IOMT platform that can detect any disease from early symptoms, LSTM will be a great help in creating the model.

8.3 CNN

A convolutional neural network (CNN/ConvNet) is the form of a deep neural network used in deep learning to analyze visual pictures. Nowadays most of the people think neural networks to be matrix multiplications, but this is not the case with ConvNet. It uses a process known as convolution. A third function that specifies how the shape of one is changed by the other is produced by performing the mathematical process of convolution on two functions.

9. Conclusion

In this study, we review several papers and try out four approaches on a dataset to see how accurate they are. To diagnose the disease, data were acquired from the internet. The data for this study came from the OHAS Dataset. There are 2129 rows of patient data in this csv file. We primarily used three procedures to represent the data: Data Cleaning, Data Scaling, and Redundancy Removal. We worked with a dataset that includes a variety of phases that culminate in 148 illnesses, in light of the 2129 documents in this collection, including various patient samples. The illness prediction system employs five classifiers to calculate data. There was XGBoost, Decision Tree, K-Nearest Neighbor, Logistic Regression, and SVM, to mention a few. After preprocessing and cleaning the data, we loaded the required libraries and applied our dataset in different classifiers. We discovered that the accuracy level of the implemented classifier was 86.24 percent for XGBoost, 70.276 percent for KNN, and 58.164 percent for Logistic Regression.

References

Singh, R. P., Javaid, M., Haleem, A., Vaishya, R., & Ali, S. (2020). Internet of Medical Things (IoMT) for orthopaedic in COVID-19 pandemic: Roles, challenges, and applications. Journal of Clinical Orthopaedicsand, Trauma 11(4), 713-717.

Bansal, D., Chhikara, R., Khanna, K., & Gupta, P. (2018). Comparative analysis of various machine learning algorithms for detecting dementia. Procedia computer science, 132, 1497-1502.

Yede, N., Koul, R., Harde, C., Gaurav, K., & Pagar, C. S. (2021). General disease prediction based on symptoms provided by patient. Open Access International Journal of Science & Engineering, 6.

Joyia, G. J., Liaqat, R. M., Farooq, A., & Rehman, S. (2017). Internet of medical things (IoMT): Applications, benefits and future challenges in healthcare

domain. J. Commun., 12(4), 240-247.

Khan, M. A., & Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access, 8, 122259-122269.

So, A., Hooshyar, D., Park, K. W., & Lim, H. S. (2017). Early diagnosis of dementia from clinical data by machine learning techniques. Applied Sciences, 7(7), 651.

da Costa, T. D., Vara, M. D. F. F., Cristino, C. S., Zanella, T. Z., Neto, G. N. N., & Nohama, P. (2019). Breathing monitoring and pattern recognition with wearable sensors. In Wearable Devices-the Big Wave of Innovation. IntechOpen.

Kim, K. J., Hassan, M. M., Na, S. H., & Huh, E. N. (2009, December). Dementia wandering detection and activity recognition algorithm using tri-axial accelerometer sensors. In Proceedings of the 4th International Conference on Ubiquitous Information Technologies & Applications (pp. 1-5). IEEE.

Radhika, S., Shree, S. R., Divyadharsini, V. R., & Ranjitha, A. (2020). Symptoms based disease prediction using decision tree and electronic health record analysis. European Journal of Molecular & Clinical Medicine, 7(4), 2060-2066.

Ishii, H., Kimino, K., Aljehani, M., Ohe, N., & Inoue, M. (2016). An early detection system for dementia using the M2 M/IoT platform. Procedia Computer Science, 96, 1332-1340.

Ferdousi, R., Hossain, M. A., & El Saddik, A. (2021). Early-stage risk prediction of non-communicable disease using machine learning in health CPS. IEEE Access, 9, 96823-96837.

Bhanuteja, T., Kumar, K. V. N., Poornachand, K. S., Ashish, C., & Anudeep, P. Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.

Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

Rachakonda, L., Sundaravadivel, P., Mohanty, S. P., Kougianos, E., & Ganapathiraju, M. (2018, December). A smart sensor in the iomt for stress

level detection. In 2018 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS) (pp. 141-145). IEEE.

Khan, M. F., Ghazal, T. M., Said, R. A., Fatima, A., Abbas, S., Khan, M. A., ... & Khan, M. A. (2021). An iomt-enabled smart healthcare model to monitor elderly people using machine learning technique. Computational Intelligence and Neuroscience, 2021.

Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., ... & Mehendale, N. (2020). Disease prediction from various symptoms using machine learning. Available at SSRN 3661426.

Hamsagayathri, P., & Vigneshwaran, S. (2021, February). Symptoms Based Disease Prediction Using Machine Learning Techniques. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 747-752). IEEE.

Khan, M. A., & Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access, 8, 122259-122269.

Hashem, M., Vellappally, S., Fouad, H., Luqman, M., & Youssef, A. E. (2020). Predicting neurological disorders linked to oral cavity manifestations using an IoMT-based optimized neural networks. IEEE Access, 8, 190722-190733.

Rachakonda, L., Mohanty, S. P., Kougianos, E., & Sundaravadivel, P. (2019). Stress-lysis: A DNN-integrated edge device for stress level detection in the IoMT. IEEE Transactions on Consumer Electronics, 65(4), 474-483.

Guo, C., Zhang, J., Liu, Y., Xie, Y., Han, Z., & Yu, J. (2020). Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. IEEE Access, 8, 59247-59256.

Grampurohit, S., & Sagarnal, C. (2020, June). Disease prediction using machine learning algorithms. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.

Zhang, P., Huang, X., & Li, M. (2019). Disease prediction and early intervention system based on symptom similarity analysis. IEEE Access, 7, 176484-176494.

OHAS:https://www.kaggle.com/code/kerneler/starter-disease-prediction-through-112c9798-a/data?select=OHAS+Dataset.csv