How to Cite:
Choubey, R., & Gautam, P. (2022). Supervised ensemble classifier algorithm for prediction of liver disease, lung cancer and brain stroke. *International Journal of Health Sciences*, *6*(S4), 9581–9592. https://doi.org/10.53730/ijhs.v6nS4.11241

Supervised ensemble classifier algorithm for prediction of liver disease, lung cancer and brain stroke

Ravi Choubey
Rabindranath Tagore University, Raisen, India
Corresponding author email: ravichoubey808@gmail.com

Dr.Pratima Gautam
Department of CS & I.T Rabindranath Tagore University, Raisen, India

***Abstract*---**Many diseases are increasing day by day and it takes too much time to detect. In India after Covid-19 pandemic so many diseases have been spread their era. Like Liver Disease, Lung cancer and Brain Stroke. They are among us and lethal diseases which need to predict earlier or in initial stage. Machine Learning (ML) is the subset of Artificial intelligent which can imitate like human intelligence and it can process the large information. The classification or prediction of those diseases can be done by classifiers. The disease prediction is the method which can predict future of Liver diseases, Lung Cancer and Brain Stroke possibilities based on the collection of historical dataset. In this paper we will use Hybrid Ensemble Classifier Model (HECM) which is the combination of Supervised Classifiers like LightGBM, Random Forest, KNN used as Ensemble Classifier then output given to Voting classifier for final output. Accuracy and time will be calculate

***Keywords*---**liver, lung cancer, brain stroke, ensemble classification, hybrid machine learning supervised classifiers.

Introduction

Machine learning is part of data science that provides computers the ability to learn automatically and improve from experience without being explicitly programmed. Machine Learning focuses on the developments of system programs to access data and make prediction for future decision. Extracting knowledge from enormous amount of data and translating unprocessed data into valuable information called data mining. It is part of machine learning and have potential to give better result. In this technology it provides support in the

recognition of patterns among data. The various application fields in which data mining techniques are used extensively. Presently alone single classifier is not enough to classify with higher accuracy and less time. So we can build hybrid classifier using of many classifiers. This combined technique is called Hybrid Machine Learning Model. Existing research study shows comparison of many classifiers are used for Disease prediction. But they did not proposed any common model for many diseases. In this paper we introduced Hybrid Ensemble Common Model (HECM) for Disease prediction.

Disease Description

Liver Disease, is inflammation of liver cause by any foreign substance, i.e. bacteria, protozoa, virus, alcohol or any other toxic material which lead liver malfunctioning. According to WHO 2010 report so that 8.6 million people dead from liver damage or liver diseases. After next 15 years it will be increase 84 million people. The liver cancer cause by any foreign substance and also lead malfunctioning of liver cells and can occur several liver diseases that are liver cancer, hepatitis, fatty liver, Wilson, chronic liver disease. Symptoms of liver disease are abdominal pain, weight loss, yellow skin and nails, Leg swelling etc.

Stroke: A stroke happens when blood clots stops the blood flow to the brain. In this condition brain didn't get nutrition and oxygen as required. Then this can lead brain cells die. If immediately the medical assistance provided then permanent damaged or death can be cured. Sometime over bleeding, blood veins burs can cause brain stroke and it can be permanent or temporary. Smoking, heart disease, high sugar and high LDL level can cause brain stroke. Symptoms are generally, loss of memory, sudden fall, paralysis, nose bleed and headache some time brain death. As per W.H.O the brain strokes are second most leading cause of death in the whole world, 15 million people suffering from stroke worldwide and common in India

Lung Cancer, The Lung Cancer is the most deadly disease in human life. Lung Cancer (LC) cause major deaths due to late diagnosis and it has affected a number of people worldwide.[1] Lungs are responsible for exchange oxygen and carbon dioxide between human body and environment. Lung's structure are like spongy and can easily trap oxygen from environment. Cancer is Condition where Lung's Cells grow uncontrollable. In recent times, Lung cancer becomes the major reason of death. Among people of any age groups. Due to today's unhealthy life style, smoking and pollution all are main reason for Lung Cancer. Therefore, the improvement in predicting the Lung Cancer is required in the health sector with the help of different ML methods. The commonly seen symptoms of Lung Cancer are breath shortness. A new cough that doesn't go away. Chest pain, Coughing up blood, Hoarseness, Shortness of breath.

Problem Definition

The problem of disease prediction is to predict that the person or patient have disease or not. If patient have disease then it classifies as Target =1 (suffering from disease) else Target = 0 (Healthy). It classifies a disease with higher accuracy and with single model.

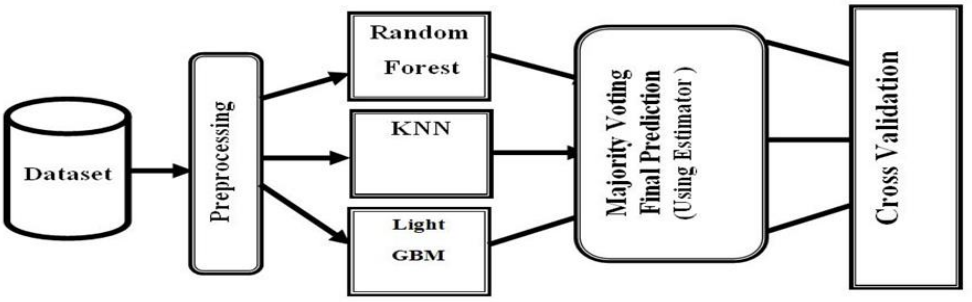
Method & Research Methodology

Qing Wu et al. (2017) In this paper they proposed novel neural network based algorithm which called Entropy Degradation Method with vectorized histogram features to detect small cell lung cancer.[2] They clustered images on the basis of histogram of image and classified that image as it is cancer positive or negative. In this study the EDM algorithm achieved accuracy 77.80% accuracy at earlier time of cancer. EDM treated as weak classifier and Adaboost used as base classifier in this study.

In this research work we mainly focused on prediction of the 3 main diseases which are Liver, Lung Cancer And Brain Stroke With the help of the Hybrid Ensemble Common Model (HECM). That model based on combining KNN, LGBM (Light Gradient Boost) and Random Forest and Estimated with Voting classification for final result. Voting classifier worked as Meta classifier and KNN, LGBM and RF classifier will use as the base classifier and Voting Classifier used as met classifier. Dataset collected from UCI Machine Learning Repository, different dataset taken for Liver, Lung Cancer And Brain Stroke. All the dataset has different features and attributes that's why we created common hybrid model for predict with high accuracy and efficiency with less time

Proposed methodology Algorithm

Following are the various steps of research methodology:-
Step 1: Input the dataset to prediction any disease. In our case we used Liver, Lung Cancer And Brain Stroke dataset.
Step 2: Preprocess dataset, remove outliers, change categorical attribute into continuous variables .Split dataset 75:25 ratio, where 75 is Training dataset and 25 is test data set .
Step 3: Append classifiers on Estimator with KNN, LGBM and Random Forest combined for the prediction of the dataset, in our case Random Forest Estimator was 100 in variable. KNN has leaf size = 4, metric='minkowski', n, n_neighbors=3. And output was collected in Estimator.
Step 4: Now predict final result with majority Voting classifier with Estimator. It predicts final result on test dataset and return variable as predicted values.
Step 5: Cross validation done after final result at last for validate our model accuracy.
Step6: Confusion metric, Sensitivity, F1, Recall and Time taken for final prediction recorded as per our requirement.



In classification, an input is consisting of the K-nearest training examples in feature space and on the other hands, the output depends upon n classification category. The output is belongs to classification category when, K-Nearest Neighbor is one in every of the main Machine Learning algorithms supported Supervised Learning technique .K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that's most almost just like the available categories. It is also called a lazy learner algorithm because it doesn't learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that's much almost just like the new data.

$$p(a) = \frac{k/n}{V}$$

LightGBM this classifier is based on tree learning algorithm, which is a gradient boosting framework. It perform distributed and efficient faster speed of training, less use of memory, higher efficiency, better accuracy. It supports parallel and GPU learning too. Light GBM perform increment of the tree vertically while other algorithm perform the increment the trees horizontally which means it (Light GBM) is tends the tree leaf-wise while other algorithm tends it level-wise. When increasing the identical leaf, Leaf-wise algorithm is able to reduce more loss in place of level-wise algorithm. There is given the explanation of the implementation of GBM and other boosting algorithms.

Random Forest, (RF), It is a supervised Machine Learning algorithm and it is perform as a bagging algorithm. In early Random Forest (RF) uses Classification And Regression Tree (CART) decision tree as weak learner which was based on the GINI coefficient to select features when we generate tree. The selected features are randomly selected, so we get random result and due to randomness it is useful to variance of model. RF do not need additional pruning and have anti over fitting.

Tool Descriptions

1. Accuracy: It is most important tool for classifier which is based on confusion matrix. It shows final accuracy on the basis of following formula.

$$\text{Accuracy} = (TP + TN) / (TP+FP+FN+TN)$$

Where :
TP - True Positive
TN - True Negative
FP - False Positive
FN – False Negative

2. Recall : After the accuracy it is the second most important tool which is the ratio of number of times the model predicts positive cases correctly to the total number of actual positive cases is known as recall.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

F1= 2*((precision*recall)/(precision+recall)

3. Precision: The ratio of number of times the model correctly predicts positive cases to the total number of positive cases predicted by it is called precision.

$$Precision = TP/(TP + FP)$$

Sensitivity/Recall:- The Sensitivity is a measurement tool which is also known as TPR (True Positive Rate) or recall.

$$\text{Sensitivity} \propto 1/\ \text{False Negative}$$

If the sum of sensitivity (TPR) and FNR would be = 1

$$TPR + FNR = 1 \qquad \ldots.(1)$$

Mathematically sensitivity calculated

$$\text{Sensitivity} = (TP) / (TP+FN) \ldots.(2)$$

True Positive:- True Positive refers that a person predicted Unhealthy and actually suffering from predicted disease.
False Positive:- False Position refers that a person predicted unhealthy, but that person is healthy.

Result and Discussion

In this work we used Anaconda's Spyder 4 which based on Python interpreter. Python is an open source programming language, which have huge Machine learning Library. It has large machine learning libraries like pandas, numpy and sklearn. Datasets are collected from the UCI machine learning repository. In this approach dataset is given to KNN, LGBM and RF in form of estimator [ ]*20

Hybrid Ensemble Common Model (HECM) as input, and this generate an output which give as input to voting classifier which is applied for the final output. In this proposed method HECM classification method which is the combination of KNN, LGBM and RF. This HECM method will reduces the time complexity of the model and increases accuracy. We worked mainly on 3 diseases e.g. Liver, Lung Cancer and Brain Stroke which description is following:

Dataset

Table 1: showing Liver Disease Dataset

Liver Disease Dataset	
• AGE :	Age of the patient
• GENDER :	Gender of the patient
• TSB :	Total Bilirubin in milligrams per deciliter (mg/dL)
• DB :	Direct Bilirubin in milligrams per deciliter (mg/dL)
• ALP :	Alkaline Phosphatase in units per liter (U/L)
• ALT (U/L):	Alanine Aminotransferase in units per liter (U/L) aka SGPT
• AST :	Aspartate Aminotransferase in units per liter (U/L) aka SGOT
• TP :	Total Proteins in grams per deciliter (g/dL)
• ALB :	Albumin in grams per deciliter (g/dL)
• AGR :	Albumin and Globulin Ratio
• Target:	Patient (patient with liver disease [1], or no disease [0])

Liver Disease: The Liver Dataset is taken from UCI Machine Learning Repository, it has 11 attributes and 410 records. 292 are liver patient and 118 are healthy persons. Target attribute show a person has disease or not and it contains 0 = Healthy or 1 = unhealthy. 310 records belong to male and 100 females. Following Table 1 is showing details attributes and description of dataset.

Lung Cancer: The Lung Cancer dataset is taken from UCI Machine Learning Repository, it has 16 attributes and 309 records. Lung Cancer patients are 270 and 39 are healthy persons. Target attribute show a person who has disease or not and it contains 0 = Healthy or 1 = unhealthy. 162 records belong to male and 147 females. Following Table 2 is showing details attributes and description of dataset:

Table 2: Showing Lung Cancer Attributes

Lung Cancer Dataset	
Gender	Male [0] Female [1]
Age	Age in integer
Smoking	Smoking Yes [2] or No[1]
Yellow Finger	Finger Coloured Yellow Yes [2] or No[1]
Anxienty	Yellow Yes [2] or No[1]
Peer Pressure	Peer Pesure Feel Yes [2] or No[1]
Chronic Disease	Any Disease Yes [2] or No[1]
Fatigue	Fatigue Yes [2] or No[1]
Allergy	Yes [2] or No[1]
Wheezing	Yes [2] or No[1]
Alcohol Cons.	Yes [2] or No[1]
Coughing	Yes [2] or No[1]
Shortness of Breath	Yes [2] or No[1]
Swallowing Difficulty	Yes [2] or No[1]
Chest Pain	Yes [2] or No[1]
Target	Patient have Disease [Unhealthy (1), Healthy (0)]

Brain Stroke: Data set is taken from Kaggle and it has 8 attributes, the dataset was very high unbalanced and had missing values and outliers which is removed during data preprocessing, the following table shows attributes and description of dataset.

Table 3:Brain Stroke Dataset

Brain Stroke Dataset	
gender	Male 0 Female 1
age	Age of Patients
hypertension	Hypertension yes 1 no 0
heart_disease	Heart Disease yes 1 no 0
avg_glucose_level	Glucose level in blood
bmi	Calculated Body mass index
smoking_status	Smoking Yes 1, No 0
stroke	Prediction of Target (yes=1), (No=0)

For each dataset classification following steps are :
Step 1: Input the dataset of any disease. Then follows next
Step 2: Preprocess of dataset, remove outliers, change categorical attribute into continuous variables. Fill NAN values = 0
Step 3: Now split the dataset 75:25 ratio, where 75 is Training dataset and 25 is test data set .
Step 4: Append all three classifiers on Estimator with KNN, LGBM and Random Forest combined for the prediction of the dataset, KNN has leaf_size=4, metric='minkowski',n, n_neighbors=3. In our approach.
Step 5: Predict final result with majority Voting classifier with Estimator. This will predict final result on test dataset and return variable as predicted values.

Step 5: The cross validation done after final result.
Step 6: Confusion metric, Sensitivity, F1 and Time taken for final prediction recorded as per our requirement. And Cross validation result printed at the end.
Step 7: For every dataset repeat step 2 to 5.

Table 4: Comparison of Liver Disease models with our proposed model

	Author	Model	Accuracy	Precision	F 1-measure	Sensitivity /Recall	Time Second
Liver Disease	Hartatik	Naïve Bayes	72.5	70	70	70	NA
	Hartatik	K.N.N	63.19	60	60	60	06
	Test	Random Forest	79.9	78	75	72	07
	Proposed	HECM	81.20	90.91	87.91	85.11	6.4 second

In this table 4, represent comparison of several Classifiers with our proposed HECM method. In this comparison the Liver disease accuracy is compared with different machine learning algorithms. This table shows that the naïve byes accuracy 72.5% and test model Random Forest accuracy is 79.90% only. Whereas our model HECM showed 81.20% accuracy. Precision, F1 and Sensitivity is higher among all classifiers. Time is also calculated and it shown 6 second.

Figure 1: The Line Graph plotted between accuracy and other measuring tools for Liver Disease

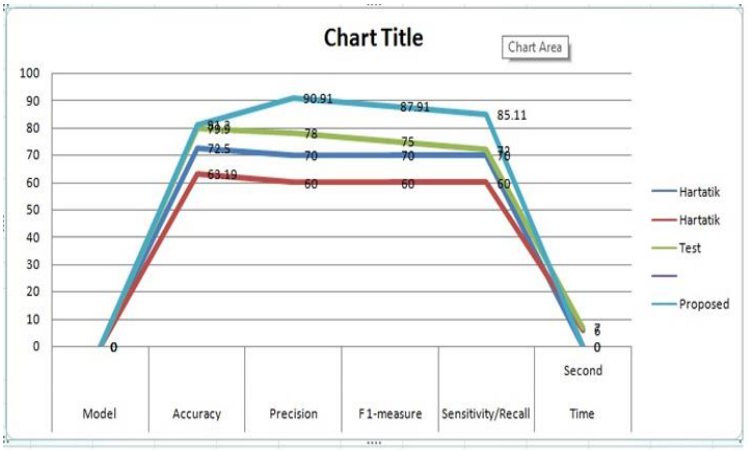


Table 5: Comparison of Lung Cancer Disease models with our proposed model

	Author	Model	Accuracy	Precision	F measure	Sensitivity/Recall	Time
Lung Cancer	Qing Wu	Neural Network algorithm	77.8	62	75	75	NA
	Radhanath Patra	RBF, KNN, ANN	81.25	77	80	81	NA
	Test	XGboost	80	79	79	79	5.2
	Proposed	HECM	90.32	94.44	94.44	94.44	3.6

In this table shows Lung cancer text dataset classification, we also mentioned author and previous works results. The table shows that previous study shown higher accuracy was 81.25 percent. Sensitivity, Precision, F1 was higher as 94.44 percent. This study shown better result than existing or previous study. The time for final prediction is 3.60 second. In figure 2 shown comparative bar graph chart of various tools and time.

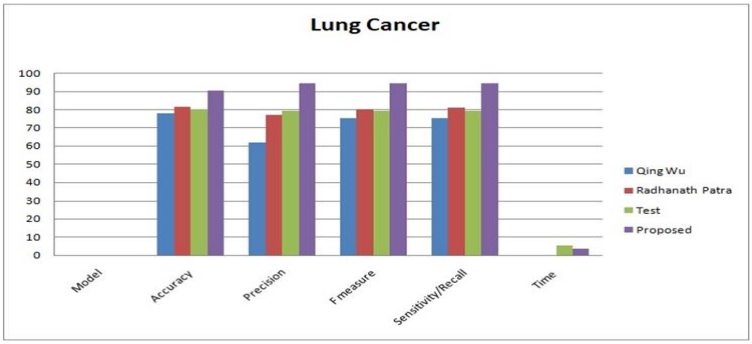


Figure 2: The bar graph between various measuring tools for Lung cancer dataset

Table 6: Comparison of Brain Stroke models with our proposed model

	Author	Model	Accuracy	Precision	F 1-measure	Sensitivity /Recall	Time Second
Brain Stroke	Soumyabrata Dev	Neural Network	78	80`	75	70	NA
	Test	SVM	88	80	81	81	05
	Test	Random Forest	91	88	88	70	7.52
	Proposed	HECM	98.08	100	76.92	62.50	5.34

In this table 6 we compared 3 model or classifier with our model HECM. Soumybrata Dev et al. did worked on Neural Network [9] achieved 78 percent accuracy with 80 % other parameters. Random forest shows 91% accuracy, F1 was 88 percent and sensitivity was just 81.00 % while our proposed model HECM showed 98.08% accuracy with 98% cross validation and 100% with precision. Brin Stroke prediction got the best result with our model. And time for final prediction was just 5.34 seconds. In figure 3 the comparison shown by line graph.

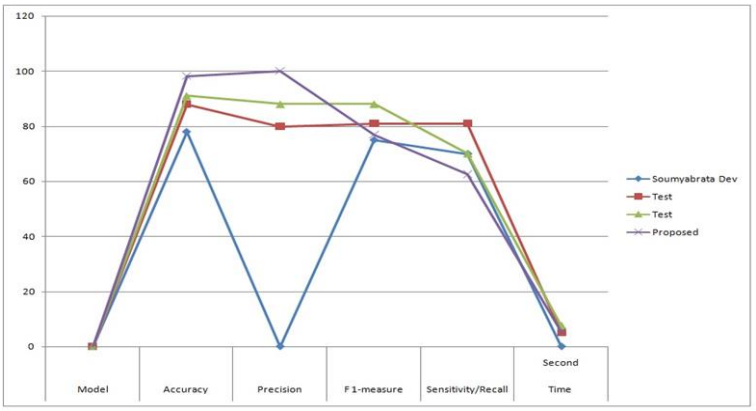


Figure 3: The line graph between various measuring tools for Brain Stroke Disease

Conclusion

In this study, we introduced a Hybrid Ensemble Common Model (HECM) which was based on supervised classifiers e.g. KNN, Light GBM and Random Forest also as Estimator or final classification done by soft Voting. In previous research work they used combination of two classifiers or single classifier and they also did not recorded the time for the classification. The average accuracy was just 84% in previous work which was good but still need higher accuracy and need single model for classification of almost all diseases. While real time frame work or working with web based classification, we need more accuracy and low time consumption. In this proposed HECM model we achieved 98 per cent accuracy with Brain Stroke dataset, and 100% Precision. For Lung Cancer dataset achieved 90.32 per cent accuracy. For Liver disease we got 81.20 per cent accuracy which was good. All diseases achieved higher accuracy and less time consumption than previous study.

Acknowledgments

We would also like to express our gratitude to all the my contributors, namely the co authors, reviewers, and editors, who have made this issue possible. IJHS is currently accepting manuscripts for upcoming issues based on original qualitative research that opens new areas of inquiry and investigation to all researcher.

References

1. Choubey, Ravi & Gautam, Pratima. (2021). Analysis of various machine learning Algorithms for Heart Disease Prediction. PIMT 13. 4.
2. D. Reddy, E. N. Hemanth Kumar, D. Reddy and M. P, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method," 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 353-357,
3. H. Hartatik, M. B. Tamam and A. Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2020, pp. 1-5, doi: 10.1109/ICORIS50180.2020.9320797.
4. H. Hartatik, M. B. Tamam and A. Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2020, pp. 1-5, doi: 10.1109/ICORIS50180.2020.9320797.
5. https://doi.org/10.1016/j.lungcan.2021.01.07.
6. Khikmatullaeva, Khaydarov, N. K., Abdullaeva, M. B., & Aktamova, M. U. (2021). Cognitive disorders in stroke. *International Journal of Health & Medical Sciences*, *4*(2), 202-207. https://doi.org/10.31295/ijhms.v4n2.1700
7. M. I. Faisal, S. Bashir, Z. S. Khan and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018, pp. 1-4, doi: 10.1109/ICEEST.2018.8643311.
8. M. Selma, A. Mohamed, H. M. Yassine and B. Issam, "How to have a structured database for lung cancer segmentation using deep learning technologies," 2021 International Conference on Networking and Advanced Systems (ICNAS), 2021, pp. 1-5, doi: 10.1109/ICNAS53565.2021.9628946.
9. Marjolein A. et al. “Lung cancer prediction by Deep Learning to identify benign lung nodules,” Volume 154, 2021,
10. N. Afreen, R. Patel, M. Ahmed and M. Sameer, "A Novel Machine Learning Approach Using Boosting Algorithm for Liver Disease Classification," *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702488.
11. O. Günaydin, M. Günay and Ö. Şengel, "Comparison of Lung Cancer Detection Algorithms," 2019 Scientific Meeting on Electrical-Electronics&Biomedical Engineering and Computer Science (EBBT), 2019,pp.1-4**.**
12. Patra R. (2020) Prediction of Lung Cancer Using Machine Learning Classifier. In: Chaubey N., Parikh S., Amin K. (eds) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore.
13. Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC*)*, 2017, pp. 88-91, doi: 10.1109/ISCSIC.2017.22
14. R. P.R., R. A. S. Nair and V. G., "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," 2019 IEEE International

Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-4.
15. Soumyabrata Dev et al. “A predictive analytics approach for stroke prediction using machine learning and neural networks”, Healthcare Analytics,Volume 2,2022, 100032,ISSN 2772-4425,
16. Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Get vaccinated when it is your turn and follow the local guidelines. *International Journal of Health Sciences*, *5*(3), x-xv. https://doi.org/10.53730/ijhs.v5n3.2938