www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

Study of Interpretability in ML Algorithms for Disease Prognosis

P. Archana Menon[1]; Dr.R. Gunasundari[2]
[1]Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India.
[1]archananirmal1414@gmail.com
[2]Professor & Head, Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, India.
[2]gunasoundar04@gmail.com

Abstract
Disease prognosis plays an important role in healthcare. Diagnosing disease at an early stage is crucial to provide treatment to the patient at the earliest in order to save his/her life or to at least reduce the severity of the disease. Application of Machine Learning algorithms is a promising area for the early and accurate diagnosis of chronic diseases. The black-box approach of Machine Learning models has been circumvented by providing different Interpretability methods. The importance of interpretability in health care field especially while taking decisions in life threatening diseases is crucial. Interpretable model increases the confidence of a medical practitioner in taking decisions. This paper gives an insight to the importance of explanations as well as the interpretability methods applied to different machine learning and deep learning models developed in recent years.

Key-words: Interpretability, Explainability, Disease Prognosis, Machine Learning, Deep Learning, Blackbox, Whitebox, Visual Representation**.**

1. Introduction

Prognosis of any disease defines the estimate of the likely course and outcome of the disease. A doctor can predict the disease by considering the patient's symptoms. By applying data mining techniques and machine learning algorithms on patient's health data, a well-developed model can also predict the disease. Early detection of a disease helps to start the early treatment and hence could save many lives from life threatening diseases or at least reduce the severity of the disease. Different ML methods are applied for Cancer prognosis and also for classification of patients into low and high-risk categories.

As an effort of many years, researchers have collected information about the people with similar type of disease. With the help of these statistics, doctors can estimate disease prognosis. Because statistics are based on large groups of people, they cannot be used to predict exactly what will happen to you [4]. So we need analysing tools for medical data to diagnose, predict or classify a disease, find out its severity, recommend personalised drugs etc.

Machine learning is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [5]. The intelligent systems built on machine learning algorithms have the capability to learn from past experience or historical data [5]. Such systems are capable of performing medical diagnosis, disease prediction, image processing, disease classification, learning association, regression etc. Deep learning is a subset of machine learning in artificial intelligence concerned with algorithms inspired by the structure and function of the brain called Artificial Neural Network [6].

Machine learning interpretability is critical for data scientists and researchers to explain their models and understand the value and accuracy of their findings. Interpretability is the degree to which a model can be understood in human terms [25]. ML techniques follows Black box approach. It rarely explain the prediction in an understandable form for users [2].

In this paper, the importance of interpretability in disease prognosis and the recent investigations of interpretability in different machine learning models are discussed. The remaining sections of this paper is designed as follows: Section II says about the motivation for writing this paper. Section III covers the Interpretability in ML algorithms which discusses about the importance of Interpretability, various Interpretability methods available, different types of data used for disease prognosis, and a study of Interpretability methods chosen by different researchers for their models in recent years. Tables in Section IV shows the different ML models, Interpretability methods and performance measures of different problems. Section V concludes the study into key observations.

2. Motivation

Disease prognosis is crucial in medical decision making, and such decisions are significant to patients. Interpretability of systems is essential in the prognosis of many life threatening diseases since it is considered as a critical and a “life or death” prediction model. Prognosis in such diseases require high forecasting accuracy and interpretation which are difficult to achieve simultaneously. There is always a trade-off between interpretation and accuracy. Hence, developing an accurate and interpretable model at the same time is a very challenging task. High accuracy often requires

developing complicated black box models while interpretation requires developing simple and less complicated models, which are often less accurate [2].

Blackbox approach has become a bottleneck to Neural Network models. How to address the problem of explainability or interpretability is a big challenge for the developers. EHR, medical images and other medical databases are extremely utilized for Natural Language Processing (NLP), image processing, text mining, computer vision etc. ML and DL models are developed by making use of these processed information to successfully apply it in medicine or healthcare industries. Visualization helps both domain experts and end users with better understandability of the model. Such visualization techniques assist a doctor to find out the reasoning behind model predictions and hence could explain the same to his patients in a much better way.

Fig. 1 - Computer, Data Analyst and Medical Expert Interaction Cycle

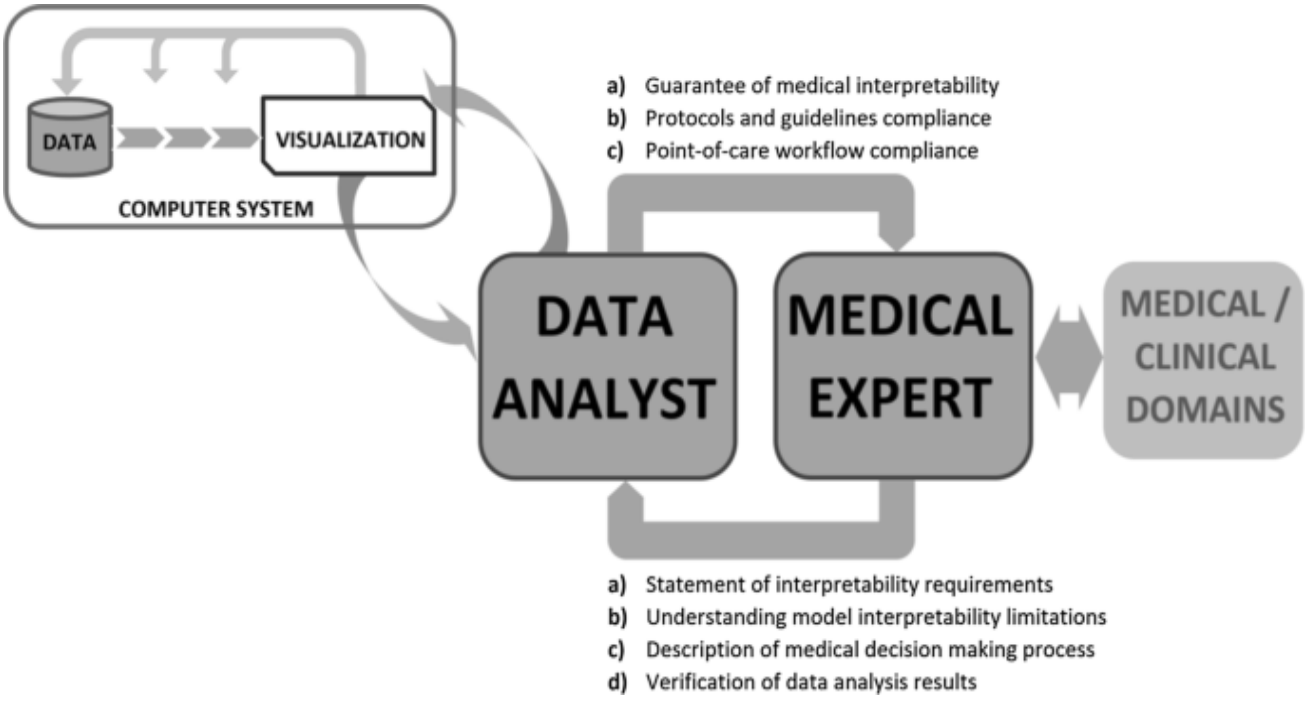


Fig. 1. [26] depicts the involvement and interaction between computer system, data analyst, and medical expert for better and accurate clinical decision making. The data and the model will be revised and evaluated by the experts. The analyst will go through certain data pre-processing stages and then modify and optimise the model. They guarantees interpretability through visualization techniques. A greater involvement of experts from medical domain is much required to develop models in biomedical field.

3. Interpretability in Machine Learning Algorithms

Importance of Interpretability

Classification accuracy is not the only factor which one can depend on for making crucial decisions. For certain problems the model must explain how or why it came to that prediction. In ML, how to measure interpretability is not well defined. Interpretability helps in finding out the bias in the models and it also increases the social acceptance of the model [7].

Explanations always help to ensure the fairness, privacy, reliability and trust of the ML models we employed. When we try to explain a model, the explanations must have some desirable properties such as it must be accurate, translucent, stable, consistent, comprehensive, and certain. More than everything, a good explanation must be human friendly [7].

Interpretability or Explainability Methods

There are several ways to achieve interpretability in our ML model. First and simplest of all, is by implementing a set of algorithms which produces interpretable results implicitly. These are called as Intrinsic methods. Linear regression, Logistic regression, Decision tree etc. are examples of such models [7].

The second way to achieve interpretability is by separating explanations from the model. Such models are called as model-agnostic interpretation methods. Here, developers can build models using any ML algorithm and then an independent interpretable technique can be applied on the model. This makes the model-agnostic interpretation methods more flexible in terms of model, explanation and representation. Partial Dependency Plot (PDP), Global Surrogate, Local Surrogate (LIME), Shapley values, SHAP etc. are examples of model-agnostic models [7].

Explaining the predictions made by Neural Network is more difficult because of the complex mathematical computations taking place inside the hidden layers of the network. More specific interpretation methods are developed for neural networks even though model-agnostic methods can still be used for neural network explanation. Some of the interpretation methods specifically used for explaining neural networks are as follows:

- Learned Features approach where Convolutional Neural Network learns abstract features from image pixels [7].
- Saliency maps highlight the pixels that were relevant for an image classification. Grad CAM, Guided Grad-CAM, SmoothGrad etc. comes under Saliency map [7].

- Concept-based approach generate explanations based on concepts such as colour, object, abstraction or an idea and are not limited to the feature space of the network [7].

Fig. 2 - Big picture of Explainable ML

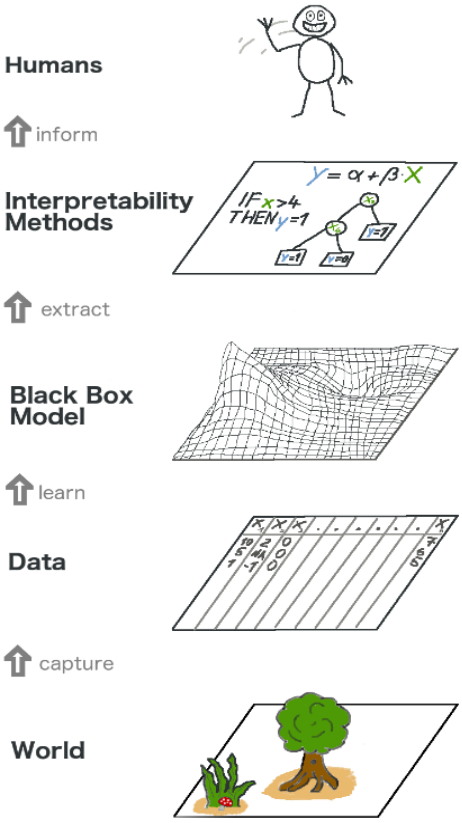


Fig. 2. [7] shows that the data from the real world, which is at the lowest level, is captured using different techniques and applied to black box ML models upon which interpretability methods are applied to give explanations to humans. These high level explanations could be in visual or textual form.

Python provides an open-source package called InterpretML [11] which provides 2 types of interpretability. One is Glassbox, in which ML models are interpretable by themselves and the other is Blackbox, in which explainability techniques are to be separately added to the ML models. Comparison of different interpretability algorithms can be carried out easily using this package.

Types of Data for Disease Prognosis

The dataset used for medical analysis could take the form of EHR, EMR, images, physiological features, tabular data etc. An EHR or Electronic Health Record is the systemized

collection of patient information and storing it in digital format. EHRs store data accurately and captures the state of a patient across time. It eliminates redundancy and the need to track down a patient's earlier medical records. It ensures that the data is up-to-date, error-free and clear. The biggest problem with EHR is the difficulty in arranging and processing high dimensional heterogeneous data for diagnosis and prediction. Also it faces privacy and cyber security issues [1]. Medical images give more and accurate information about a particular disease and it is easy to manage and retrieve useful information from images after performing a few pre-processing steps on it.

Widespread availability of medical imaging systems resulted in the emerging of medical images. By analysing these images of a patient, a domain expert can diagnose the disease and its severity. The images could be in the form of X-ray, CT, MRI, PET scan, mammography, histopathological images etc. These images are important in classifying the disease stages, clinical decision-making including therapy, follow-up, predicting the severity of the disease as well as the survival rate of patients. A lot of work had been carried out using the medical image dataset to diagnose and predict diseases with classical ML techniques and Deep Learning Techniques.

Study of Interpretability Methods used in ML Models

A detailed study is carried out regarding the ML techniques used in the prognosis of various types of diseases. The explanation methods used by each researcher for their model's interpretability is also observed. Most recent and relevant papers are included here.

The study is made on various category of diseases. Initial two paper makes use of EHR data for disease prediction. The third paper deals with molecular biology for predicting cellular responses. Next few papers are based on breast cancer, glioma cancer and lung cancer prediction. Then a study on brain MRI images for brain lesion segmentation and plaque classification in Alzhimer's disease is made. Next few papers concentrated on skin tumour classification, segmentation of colorectal polyps, and prediction of Chronic Kidney Disease. Then a study on the Classification of Diabetic Retinopathy Disease and detection of Parkinson's disease is carried out. And the last three papers are associated with COVID-19 disease detection and prediction.

In the work proposed by [8], a Dipole Bidirectional RNN was developed to remember the patient information of both past and future hospital visits and introduced three attention mechanisms to gauge the relationship of different visits for diagnosis prediction. Different models evolved before based on RNN-based approaches suffered from problems such as the performance of RNNs goes

down when the length of the sequences are huge, and the relationship between subsequent visits are not considered.

Mixing diagnosis with treatment using EHR data affect prediction results. Cross Over Attention Model (COAM) proposed by [9], uses two RNNs to separate the diagnosis information from the treatment information to ensure the integrity and uses the mutual relationship between them to improve the accuracy of the disease onset predictions. The cross over attention mechanism in COAM provides interpretable prediction results and clinically meaningful explanations.

[10] proposed a hybrid approach called CellBox that combines explicit mathematical models of molecular interactions with efficient parameter inference algorithms adapted from deep learning to predict cellular responses. The model do not require prior knowledge and are data driven. They used the dataset of a melanoma cell line and achieved global optimal in a complex multidimensional space and could interpret the solutions using ODE (Ordinary Differential Equation) applied on CellBox. The approach was readily applicable to different models of cell biology and it outperformed the previous BP dynamic model approaches.

An interpretable ML framework developed by [12] consists of a feature extraction module to extract and create transparent and meaningful high level features of images and an explanation extraction module for creating good explanations based on extracted features. Brain tumour MRI images were used for predicting glioma cancer in an interpretable and explainable manner. Model's explanation is achieved through two basic language forms- graph diagrams and questions–answers form.

Dominance Classifier and Predictor (DCP) algorithm [13] is capable of automating the process of discovering human-understandable, domain-explainable, more accurate ML models. DCP could achieve higher accuracy in its interpretable ML method on the benchmark Wisconsin Breast Cancer Dataset. DCP could provide human friendly, simple, visual explanations and textual explanations for the model.

Using radiomic features, lung cancer prediction pipeline has been designed by [14]. The SISC (Stacked Interpretable Sequencing Cells) architecture of radiomic sequencer comprises of interpretable sequencing cells. This offers prediction interpretability in the form of critical response maps. Critical maps highlights the critical regions used by the sequencer for making predictions. Critical maps validates the predictions and provides a good collaboration between radiologist and machine for effective diagnosis.

Fig. 3 - Critical response map for Benign Cases. (a) Benign Nodule from Lung CT-images and (b) corresponding Critical response map

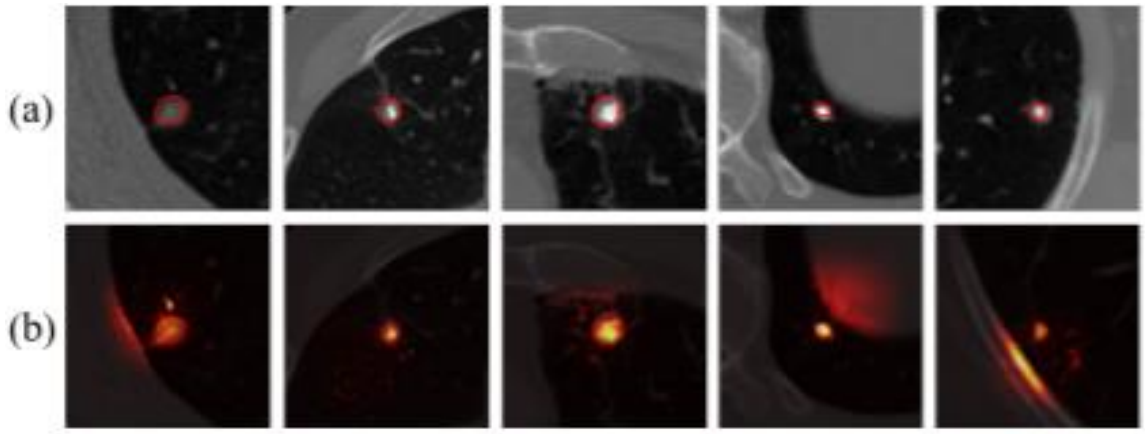


Fig. 4 - Critical response map for Malignant Cases. (a) Malignant Nodule from Lung CT-images and (b) corresponding Critical response map

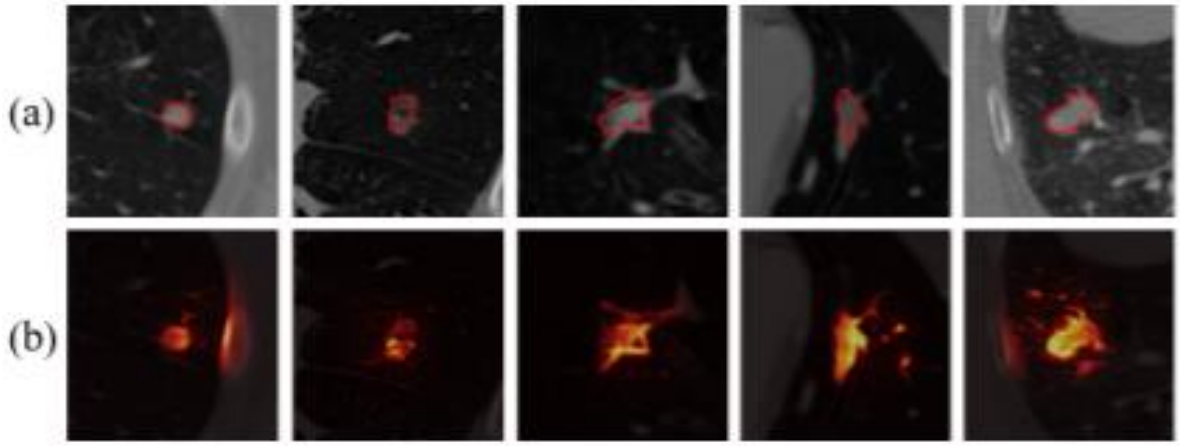


Fig. 3. and Fig. 4. [14] shows how critical response map represents critical regions while predicting whether a lung nodule is malignant or benign.

[15] developed a model for the interpretation of automatically learnt features for lesion segmentation in medical images. For unsupervised feature learning, Restricted Boltzmann Machine (RBM) is used and for classification, a Random Forest classifier is employed. The model made both global interpretations- in order to understand whether the system learnt relevant relations in the data correctly, and local interpretations- to perform predictions on a voxel and patient level. Proposed methodology was evaluated in brain tumour MRI and ischemic stroke lesions.

[16] presented neuropathological analysis of Aβ pathologies in Whole Slide Images of human brain for the classification of Alzhimer's disease. They pipelined an automated segmentation for image pre-processing, a cloud-based interface for annotation of images and CNN models for

distinguishing Aβ pathologies as cored plaques, diffuse plaques or cerebral amyloid angiopathy. Interpretability is achieved through saliency mapping which demonstrates that networks learn patterns from pathologic features.

The neural network RECOD (REasoning for Complex Data) [17] makes use of dermoscopic images to classify three skin tumours and is validated on PH^2 and derm7pt datasets. CAV (Concept Activation Vectors) and TCAV scores are used for the Concept-based explanation of the model.

An interpretable CNN model [18] based on FCN 8 (Fully Convolutional Networks) was developed for the semantic segmentation of colorectal polyps. It also shows the uncertainty in the importance of input features affecting precise predictions using Monte carlo Guided Backpropagation. Interpretability to the model is achieved through Guided Backpropagation.

Risk of developing a Chronic Kidney Disease (CKD) is predicted by a Neural Network- based classifier [19] using the patient's demographic and pathological data. Explanations to the model are proposed by Case-Based Reasoning (CBR) paradigm which gives explanation-by-example justification to a neural network's prediction.

[20] used Adaboost algorithm to integrate multiple deep learning models for automatic image-level Diabetic Retinopathy (DR) disease detection and classification. Weighted class activation maps (CAMs) demonstrated the suspected position of lesions.

[21] proposed an ML model that classifies SPECT DaTscan images as having Parkinson's disease or not along with the reason for prediction. Local Interpretable Model-agnostic Explanation (LIME) methods were used here to generate the visual reasoning.

To estimate uncertainty in Deep Learning solutions, [22] proposed a Monte-Carlo Dropweights Bayesian Convolutional Neural Network. The network helps in improving the diagnosis and treatment of Covid-19. They proved that the uncertainty in prediction model is strongly correlated with the prediction accuracy enabling the identification of false predictions or unknown cases. Uncertainty and interpretability are visualized using saliency maps generated by various state-of-the-art methods.

A deep learning model developed by [23] uses CT scan and chest X-ray images to detect COVID-19 patients. Out of eight different deep learning approaches experimented, NasNetMobile generated greater accuracy. LIME is used to explain the model's interpretability.

To predict the severity of COVID-19, [24] established an interpretable machine learning model. Using clinical information, laboratory tests, and chest CT features, XGBoost algorithm predicted the possibility of COVID-19 patients becoming severe and critically ill and output the most crucial deterioration factors. Shapley Additive explanations (SHAP) gave interpretability to the

model by computing each feature's contribution individually or combined and it could find out the factors that put the patients at risk.

4. Discussion

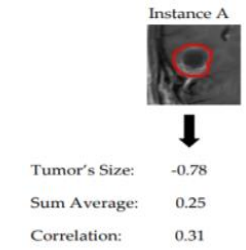
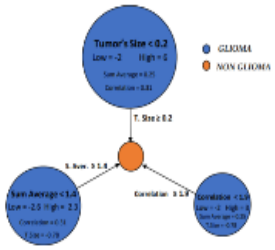
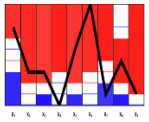
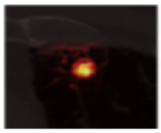
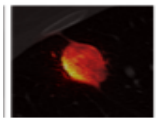
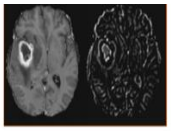
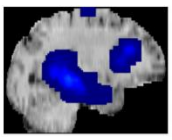
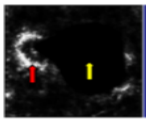
After reviewing the latest papers on different types of disease prognosis which employs new and state-of-the-art ML models, a few relevant papers are chosen here which brought novelty in their methodology and could produce improved performance measures. Table I. shows different methods or algorithms used for solving particular problems and their respective performance measures.

Table I - Methods Used and Performance Measures of Different Models

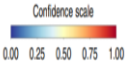
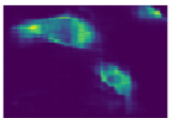
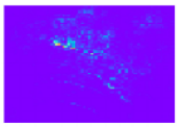
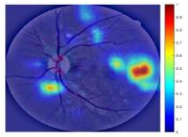
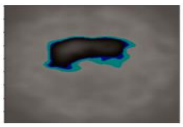
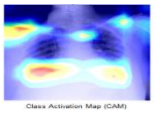
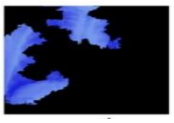
Reference Paper	Year	Problem	Methodology/ Algorithm used	Performance Measurement
[8]	2017	Disease prediction using EHR data	Bidirectional RNN using Simple Attention Mechanism	Accuracy-84%
[9]	2019	Disease prediction using EHR data	RNNs and Cross Over Attention Mechanism	Accuracy-85%
[10]	2021	Predict Cellular Responses	A Combination of Mathematical and Deep Learning Model	Highly correlated predictions
[12]	2020	Image Classification- Case study:- Glioma cancer prediction	White Box (DT, NB, LR) and Black Box (NN, SVM, KNN) methods implemented. White Box Models produced greater accuracy	AUC for LR- 94%
[13]	2018	Classifier and Predictor Algorithm- Case study:- Breast Cancer Classification	Construct Class Dominance intervals and combine it using Voting methods	Average Accuracy- 97.01%
[14]	2019	Lung Cancer Detection	Stacked CNN	AUC-89.06%
[15]	2018	Extract ML features automatically and interpretably - Case study:- Brain lesion segmentation	RBM (Restricted Boltzmann Machine) for Feature Learning & RF (Random Forest) for Classification	Accuracy of 84% in BRATS challenge
[16]	2019	Plaque classification in Alzhimer's disease	CNN pipeline	AUC for cored plaque- 74.3% AUC for diffuse plaque- 99.7%
[17]	2020	Skin tumour Classification	Transfer Learning with ensemble methods	Mean AUC - 90.8%
[18]	2019	Semantic Segmentation of Colorectal Polyps	(FCN 8) Fully Convolutional Networks, Segnet and UNet. FCN-8 outperformed others	Global Accuracy- 94.9%
[19]	2019	Chronic Kidney Disease Prediction	Neural Network	Accuracy- 95%
[20]	2019	Diabetic Retinopathy Disease Classification	DL models integrated using Adaboost algorithm	AUC- 94.6%
[21]	2020	Detection of Parkinson's disease	VGG16 CNN architecture with Transfer Learning	Accuracy- 95.2%
[22]	2020	Detecting Uncertainity in COVID-19 detection	Bayesian CNN (BCC) with Monte carlo DropWeights	Strong correlation of 0.99 between entropy of the probabilities shows uncertainty and prediction errors
[23]	2020	Detecting COVID-19 patients	8 DL techniques with transfer Learning were used. NasNetMobile outperformed others.	Accuracy in CT Scan Dataset - 82.94%. Accuracy in X-ray Dataset - 93.94%
[24]	2021	Prediction of Crucial Factors and Severity of COVID-19 Patients	XGBoost	AUC- 92.4%, Sensitivity- 90.91%, Specificity-97.96%

In this paper, since more importance is given to the Interpretability in ML algorithms, Table II. represents a summary of visual interpretations made by different models discussed in section III- D.

Table II - Representation of Different Model's Interpretabilty

Reference Paper	**Type of Data used for Interprettion**	**Interpretability Representation**	**Model Interpretability Explanation**
[12]	Brain MRI images for predicting Glioma cancer		Instance A
			Explanation graph for instance A
[13]	Wisconsin Breast Cancer Dataset for Breast Cancer Classification		DCP visual explanation. Dominant intervals of attributes in red and blue class allows direct judgement
[14]	Lung Image Dataset from LIDC- CT scan images for Lung cancer detection		Critical map for benign lung CT image
			Critical map for malignant lung CT image
[15]	Brain MRI images for Brain tumour prediction		Global Interpretability on MRI sequences
			Local Interpretability on MRI sequences
[16]	Whole Slide Image of brain images for		Heat maps for diffuse (red) and cored (yellow) plaques

Received: 15.07.2021 – Accepted: 13.08.2021

	Alzhimer's disease classification		Confidence Scale of Heat map
[18]	Colonoscopic Images for Semantic segmentation of Colorectal polyps		Uncertainty map using Montecarlo guided Backpropagation
			Interpretability map using Guided Backpropagation
[20]	Fundus images for Diabetic Retinopathy disease detection and classification		CAM of Integrated model
[21]	SPECT DaTscan images for Parkinson's disease classification		Raw Image
			With LIME
[22]	Chest X-ray images for diagonising Covid-19 disease		Saliency map- Class Activation Map (CAM)
[23]	Chest CT scan and X-ray images for identifying COVID-19 patients		LIME, applied on Chest X-ray Image of a covid-19 patient, extracting top features

5. Conclusion

Providing accurate prognostic information is the responsibility of medico-legal teams. The importance of prognosis in the field of oncology and other chronic diseases is likely to increase in the future without any doubt. Many models for disease prognosis are developing day by day. But model's Interpretability is a crucial factor for model evaluation especially in medical field. Interpretable Machine learning methods in disease prognosis based on medical images, is not a well-researched area.

In this paper, the importance of interpretability in medical field is explained along with the description of different interpretability models available. A study is made on different interpretability methods adopted by researchers in their ML models for various disease prognosis.

Improving the interpretability of ML models could not only substantially enhance the acceptability of ML predictions but also enhance the transparency of health communication between physicians and patients. Interpretable models assist doctors to recognize and appreciate the rationale behind the recommendations and helps in making accurate and confident decision.

Acknowledgment

I thank Mr. Shailesh S for the support he extended to this work.

References

Menachemi, Nir, and Taleah H Collum. *Benefits and drawbacks of electronic health record systems*", Risk management and healthcare policy vol. 4 (2011): 47-55. doi:10.2147/RMHP.S12985.

Mathew Stewart, *Guide to Interpretable Machine Learning: Techniques to dispel the black box myth of deep learning,* Mar. 2020. 2021. https://towardsdatascience.com/guide-to-interpretable-machine-learning-d40e8a64b6cf.

"*Medical Imaging*", Feb. 2018. 2021. https://en.wikipedia.org/wiki/Medical_imaging.

"*What is cancer?*". 2021https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

H. Thomas Davenport, *Evolution of Machine Learning,* 2021. https://www.sas.com/en_in/insights/analytics/machine-learning.html.

Jason Browniee, *What is Deep Learning*, Machine Learning Mastery, Aug. 2019. Accessed on: Jun. 13, 2021. https://machinelearningmastery.com/what-is-deep-learning/.

Molnar, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2019. https://christophm.github.io/interpretable-ml-book/.

Ma, Fenglong & Chitta, Radha & Zhou, Jing & You, Quanzeng & Sun, Tong. (2017). *Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks.* 10.1145/3097983.3098088.

W. Guo, W. Ge, L. Cui, H. Li and L. Kong, "An Interpretable Disease Onset Predictive Model Using Crossover Attention Mechanism from Electronic Health Records," *In IEEE Access*, 7, 134236-134244, 2019, doi: 10.1109/ACCESS.2019.2928579.

Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S. Marks, John Ingraham, Chris Sander, CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy, *Cell Systems*, 12(2), 2021, 128-140.e4, ISSN 2405-4712, https://doi.org/10.1016/j.cels.2020.11.013.

InterpretML: A Unified Framework for Machine Learning Interpretability (7) -Citation Nori, Harsha & Jenkins, Samuel & Koch, Paul & Caruana, Rich. (2019). *InterpretML: A Unified Framework for Machine Learning Interpretability.*

Pintelas, Emmanuel & Liaskos, Meletis & Livieris, Ioannis & Kotsiantis, Sotiris & Pintelas, P. (2020). Explainable Machine Learning Framework for Image Classification Problems: Case Study on Glioma Cancer Prediction. *Journal of Imaging*. 6. 10.3390/jimaging6060037.

B. Kovalerchuk and N. Neuhaus, "Toward Efficient Automation of Interpretable Machine Learning", 2018 *IEEE International Conference on Big Data (Big Data)*, 2018, 4940-4947. doi:10.1109/BigData.2018.8622433.

D. Kumar, V. Sankar, D. Clausi, G.W. Taylor and A. Wong, SISC: End-to-End Interpretable Discovery Radiomics-Driven Lung Cancer Prediction via Stacked Interpretable Sequencing Cells, *IEEE Access*, 7, 145444-145454, 2019, doi: 10.1109/ACCESS.2019.2945524.

Sérgio Pereira, Raphael Meier, Richard McKinley, Roland Wiest, Victor Alves, Carlos A. Silva, Mauricio Reyes, Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation, *Medical Image Analysis*, 44, 2018, 228-244, https://doi.org/10.1016/j.media.2017.12.009.

Tang, Ziqi & Chuang, Kangway & DeCarli, Charles & Jin, Lee-Way & Beckett, Laurel & Keiser, Michael & Dugger, Brittany. (2019). Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature Communications*. 10. 10.1038/s41467-019-10212-1.

Lucieri, Adriano & Bajwa, Muhammad Naseer & Braun, Stephan & Malik, Muhammad Imran & Dengel, Andreas & Ahmed, Sheraz. (2020). *On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors.*

M. Kristoffer & Kampffmeyer, Michael & Jenssen, Robert. (2019). Uncertainty and Interpretability in Convolutional Neural Networks for Semantic Segmentation of Colorectal Polyps. *Medical Image Analysis*. 60. 101619. 10.1016/j.media.2019.101619.

Vásquez Morales, Gabriel & Martínez Monterrubio, Sergio Mauricio & Moreno Ger, Pablo & Recio-García, Juan. (2019). Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning. *IEEE Access*, 1-1. 10.1109/ACCESS.2019.2948430.

H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma and W. Qian, "An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification," 2019 *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, 2045-2048, doi: 10.1109/EMBC.2019.8857160.

Magesh, Pavan & Myloth, Richard & Tom, Rijo. (2020). An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. *Computers in Biology and Medicine.* 126. 104041. 10.1016/j.compbiomed.2020.104041.

Ghoshal, Biraja & Tucker, Allan. (2020). Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection.

Ahsan, Md Manjurul & Gupta, Kishor Datta & Islam, Mohammad & Sen, Sajib & Rahman, Md Lutfar & Hossain, Mohammad Shakhawat. (2020). COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities. *Machine Learning and Knowledge Extraction.* 2. 10.3390/make2040027.

Zheng, Bowen & Cai, Yong & Zeng, Fengxia & Lin, Min & Zheng, Jun & Chen, Weiguo & Qin, Genggeng & Guo, Yi. (2021). An Interpretable Model-Based Prediction of Severity and Crucial

Factors in Patients with COVID-19. *BioMed Research International*. 2021. 1-9. 10.1155/2021/8840835.

O. Conor Sullivan, *Interpretability in Machine Learning*, Towards Data science, Oct.2020. Accessed on: Jun. 12, 2021. https://towardsdatascience.com/interpretability-in-machine-learning-ab0cf2e66e1

A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput & Applic 32,* 18069–18083 (2020). https://doi.org/10.1007/s00521-019-04051-w

Received: 15.07.2021 – Accepted: 13.08.2021