Evaluation of feature importance methods on real clinical data

Shadi j.Khoury*-208302463,

* Department of Biomedical engineering , Tel Aviv University –
shadij@mail.tau.ac.il

Abstract- Deep learning (DL) is a rising subject in the field of healthcare and medicine. Furthermore, to have the option to use the DL models in medical care it's basic to comprehend and decipher in reasonable terms what is influencing these models and their forecasts. understanding these impacts is called DL interpretability. DL interpretability is essential in delivering a model that clinicians can trust, and we expect to make our clinical forecast models reasonable. In this work, we look to refine the comprehension of DL interpretability when applied to clinical expectation models. While existing feature importance scores show extraordinary progress in making sense of models, we exhibit their restrictions while making sense of various datasets, particularly within sight of correlations between features. Subsequently, we standardize those scores to catch the distinction between the different score techniques and use AUC-Roc measures to catch properties anticipated from feature importance score while making sense of the data and demonstrate that there exists just a single score strategy that fulfils every one of them. The Sensitivity – Permutations analysis .we dissect this technique and exhibit its benefits.

I. INTRODUCTION

Deep learning interpretability is a subject of developing significance in the field of medical care. Interpret means to explain or to present in understandable terms. With regards to DL frameworks, interpretability is the capacity to introduce or to disclose in justifiable terms to a human. By presenting literary or visual curios that give subjective comprehension of the connection between the occurrence's parts and the model's expectation. We argue that explaining predictions is an important aspect in getting humans to trust and use machine learning effectively if the explanations are faithful and intelligible. When this model is used in high-stakes circumstances, institutions favor models which are explainable over models which might be giving relatively better accuracy.[1] A better way to say this would be that when dealing with genuine real-life problems, machine learning interpretability becomes a part of the metric for a successful and usable model. When assembling a model, it is vital to understand what it is the expected outcome what features affect the model or its forecasts. This type of assessment aids in sorting out which are the preferred models, from those that require human intervention. Often, it is important to explain the model to the institutions, particularly when utilizing the model for predicting the clinical course and patient treatments are in question. With the rising number of feature importance libraries and algorithms, it becomes difficult to conclude which algorithm or approach is best for a given case. In this study, we will discuss one such aspect of machine learning interpretability — feature importance and the various algorithms of extracting features implemented on biomedical data.[2]

II. METHODS

Feature attributions and counterfactual explanations (relating to or expressing what has not happened or is not the case.) are popular approaches to explain a DL model. The former assigns an importance score to each input feature, while the latter provides input examples with minimal changes to alter the model's predictions. To understand the difference between the algorithms we plan to normalize the features scores and analyze the difference between them, we would like to also to set a threshold on the features score depending on the relevance of that feature depending on the data set used to calculate that feature score. The normalization was done using this equation:

$$\overline{S}_i = \frac{s_I}{\max_i S}$$

Where $\overline{S}_i$ is the normalized score of features i and Si is the score of feature I.

Our feature importance methods included the most cited methods and algorithms .First we start with Sensitivity-Permutations Analysis [3], Sensitivity analysis involves systematically testing a neural network's adaptation to minor variations in input features, providing insights into the model's capabilities and discerning the influence of individual features on network predictions. By gauging the network's response to subtle changes, sensitivity analysis unveils the model's robustness, interpretability, and the relative impact of features on its decision-making process.

Secondly, we apply the Gradients – Backward propagation method, [4] Backward propagation, or backpropagation, is a fundamental process in training neural networks. It involves iteratively adjusting the model's weights by computing the gradient of the loss function with respect to each weight. This gradient information is then used to update the weights in the opposite direction of the gradient, minimizing the difference between the model's predictions and the actual targets. Importantly, if the gradients are extracted during this process, they provide valuable insights into how each feature contributes to the overall output. This analysis aids in understanding the influence of individual features on the network's predictions, enhancing our comprehension of the model's decision-making process.

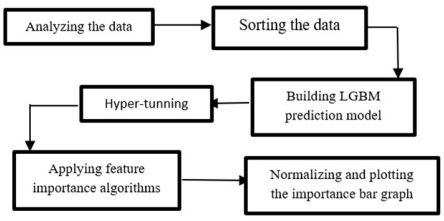
Thirdly, we apply the Activation Maximization methodology, [5] Activation maximization, often implemented through techniques like Class Activation Mapping (CAM), is a method used to visualize and understand the decision-making process of neural networks, particularly convolutional neural networks (CNNs). By maximizing the activation of specific neurons or channels in the network's layers, CAM highlights regions in the input data that strongly contribute to the network's prediction for a given class. This visualization technique not only aids in interpreting and localizing the features in the input that are most influential in driving the model's decision but also sheds light on the importance of these features in the context of deep neural networks. Activation maximization, when applied to DDNs, provides valuable insights into feature importance, offering a nuanced understanding of the specific elements that play a crucial role in the network's decision-making process during classification tasks.

Lastly, we apply Activation Maximization with Pruning , [6] Integrating Class Activation Mapping (CAM) with the strategic pruning of the most activated neurons in each layer provides dual advantages for Deep Neural Networks (DDN). Firstly, CAM enhances interpretability by visualizing critical regions in the input crucial for the network's predictions. Secondly, the concurrent pruning of highly activated neurons optimizes the network by preserving essential features while discarding less informative elements. This joint approach not only streamlines the DDN for efficiency but also facilitates nuanced feature importance extraction, offering a refined understanding of the influential factors in the decision-making process.

II.1 PREDICTION MODELS

To able to build the prediction models sorting and analyzing the features of the data is a vital point that allows us to understand what the features represent and how they affect each other by calculating correlation between different features. After analyzing the data hyper-tunning, the prediction model is important as it allows the model to get the best prediction accuracy which in turn allows the feature importance algorithms to best represent the effect of those features on the models forecast As shown in (below).

Figure 1: Schamatic showing workflow of generating the importances scores



To be able to understand the differences between different feature importance algorithms 2 different datasets were used to generate the feature importance's which will be used to compare between those algorithms.

Our deep learning networks, featuring multiple hidden layers with diverse activation functions, notably tanh and relu, a conspicuous architectural component is the inclusion of a bottleneck layer.[7] The deliberate incorporation of a bottleneck layer serves the dual purpose of information control and the regulation of generalization error within the network. This architectural choice plays a pivotal role in managing the information flow throughout the network, contributing to heightened control over model complexity and mitigating the risk of overfitting through effective regularization measures.

Moreover, the application of a bottleneck layer proves beneficial for feature extraction. By selectively allowing only the most impactful features to propagate through to the final layer, this configuration provides insights into feature importance. This focused feature propagation mechanism refines the interpretability of the network, shedding light on the specific features that exert the most substantial influence on the model's decision-making process. As can be seen in the graphs bellow.

Figure 3: Covid model NN Architectural

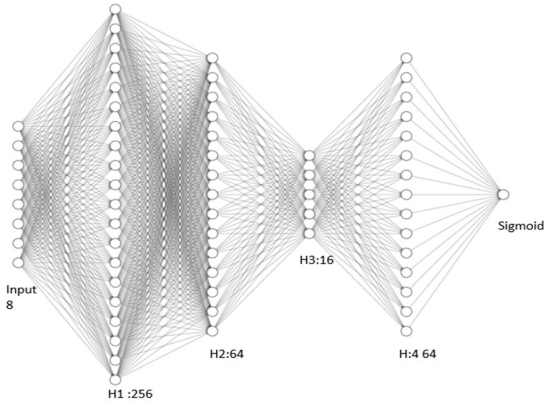
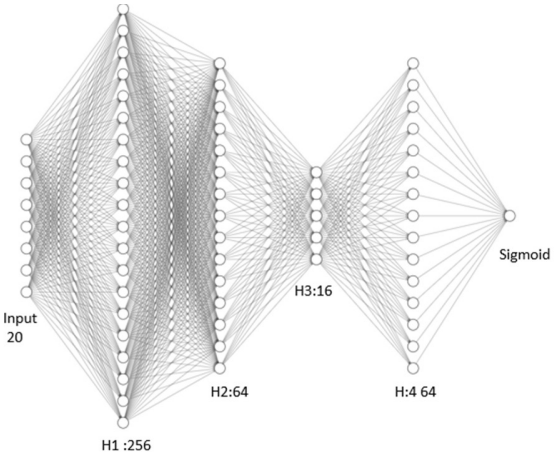


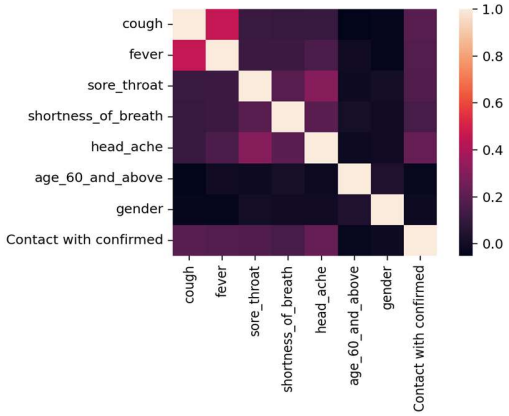
Figure 2: Bloodstream molde NN Architecture



II.2 Covid-19 Dataset

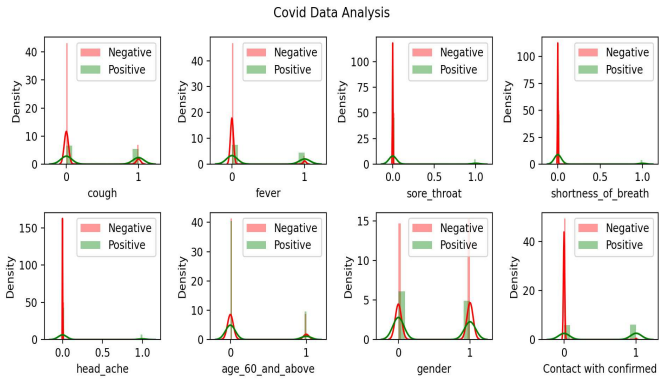
The first model was built by using covid-19 data set using only eight binary features: sex, age≥ 60 years, known contact with an infected individual...etc. On trained records from 51,831 tested individuals (of whom 4769 were confirmed to have COVID-19). While test set contained data from the subsequent week (47,401 tested individuals of whom 3624 were confirmed to have COVID-19). [8]

Figure 4: Feature vs Feature correlation heatmap for covid-19 dataset



By looking at the heatmap (*Figure 4*) there is not any significant correlation (significant meaning correlation values are higher than 70%), between the features thus proving that the features have no significant effect on each other which in turn does not affect the prediction model. To furthermore understand the relation between the features and the forecast a Probability density function (PDF) plot was generated.

Figure 5: Probability density function (PDF) of Covid features to label

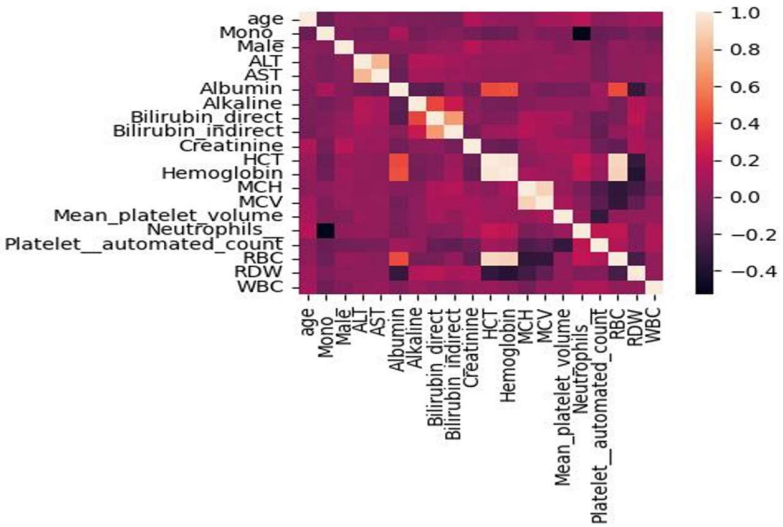


From (*Figure 5*) it is clear that most of the features point to low probability of identifying with a positive label (where positive means the individua has covid-19) ,with the exception of gender and age above 60 that in turn show high probability of identifying of positive or negative covid.

II.3 Bloodstream Dataset

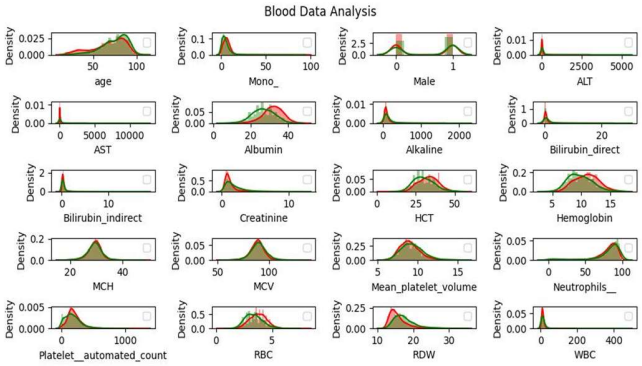
The second model was built using bloodstream infection data set using 20 features, with a mix of binary and continuous features like: age, sex, ALT, AST, Mono…etc. on trained records from that included 6,434 admissions (from years 2014-2018) and a test set that included 1,455 admissions during 2019 and the first month of 2020. [9]

Figure 6: Feature vs feature Corrlation heatmap Bloodstream data



From looking at the heatmap (*Figure 6*) it's clear that some of the feature have a significant correlation (significant meaning correlation values are bigger than 70%), between the features thus proving that the features have significant effect on each other that in turn can affect the model's accuracy. [10] To furthermore understand the relation between the features and the forecast a Probability density function (PDF) plot was generated.

Figure 7:Probability density function (PDF) of Bloodstream features to label

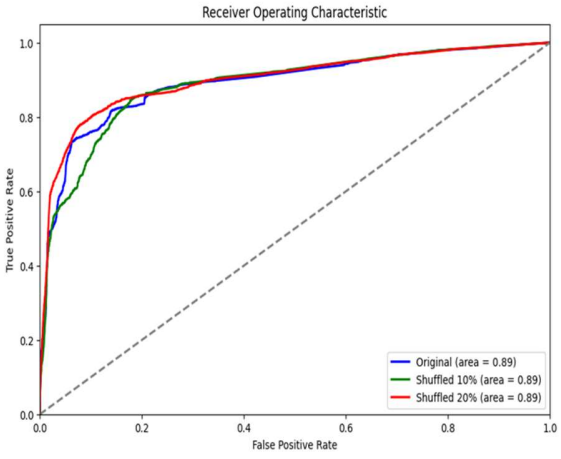


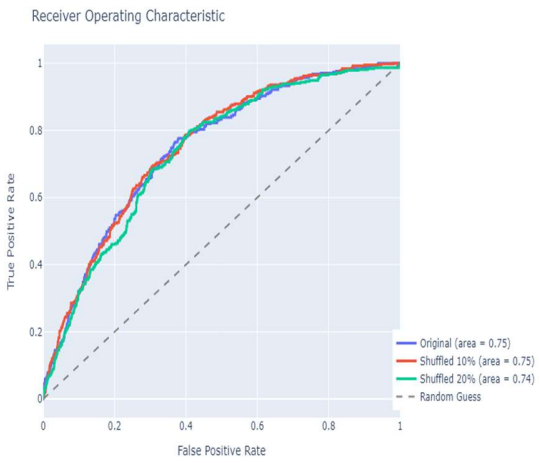
From (*Figure 7*) it's clear that most of the features point to high probability of identifying with a green label (where positive means the individua has Bloodstream infection). In turn showing that the features have a high effect on the model's forecast.

II.4 Prediction Evaluation

To evaluate the calibration of our model's predictions, we employed a Calibration method. This involved conducting multiple iterations of the training process, each time introducing slight variations to the dataset. Specifically, we considered three scenarios: firstly, the use of the original, unaltered dataset; secondly, the random shuffling of 10% of the data; and thirdly, the random shuffling of 20% of the data. Subsequently, we plotted the Area Under the Receiver Operating Characteristic (AUC_ROC) curve for each scenario. This visual representation allowed us to assess the calibration performance of our model under different data perturbation conditions. The comparison of AUC_ROC curves across these scenarios provided valuable insights into the model's ability to maintain calibration accuracy amidst variations in the input data, thereby offering a comprehensive evaluation of its robustness and reliability. As can be seen below.

Figure 8: ROC Covid-Bloodstream Predction with shffuled data





As can be seen in (*Figure 8*) in both the covid and the bloodstream model the model reached a static Auc-Roc measure. Where the COVID model is showing around 0.89 AUC and Bloodstream model is giving 0.75 AUC. Considering these AUC scores are higher than .50 means that the models can identify relationships between features and label and its not a random guess. The static AUC measures indicate that, the models consistent predictive performance in the face of minor variations in training data indicates robustness. This resilience is a positive trait, showcasing the model's reliable calibration and enhancing its generalization capability. It suggests the model has learned stable and meaningful patterns from the training data, contributing to consistent accuracy across different conditions.

III. RESULTS

III.1 Covid-19 model.

After running covid-19 data set in the model a feature importance plot was produced as can be seen in the figure below.

Figure 10: Normalized importance score's covid features Sensitivity.

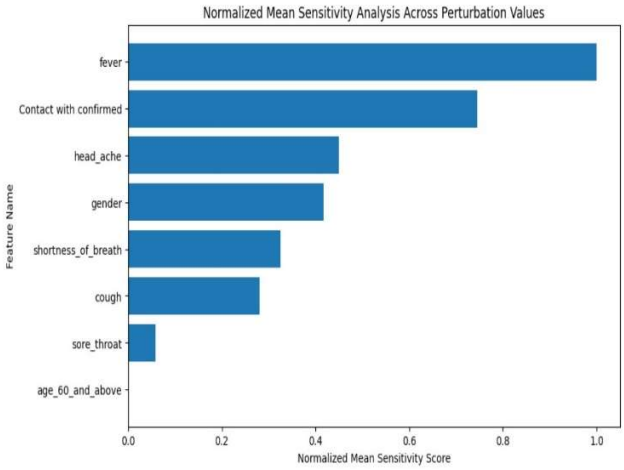


Figure 9: Normlized importnace score's covid featuers Gradints

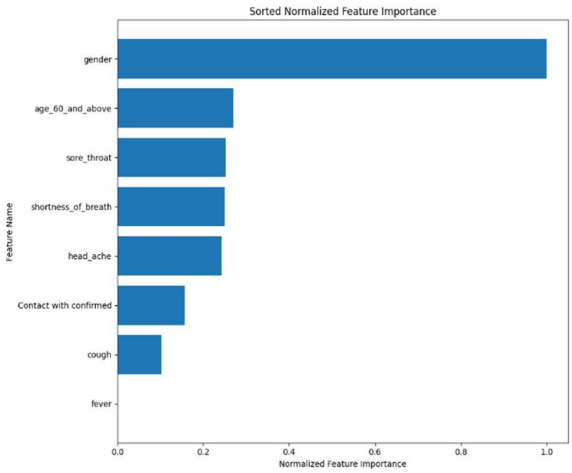


Figure 11: Normalized importance score's covid features Activation .

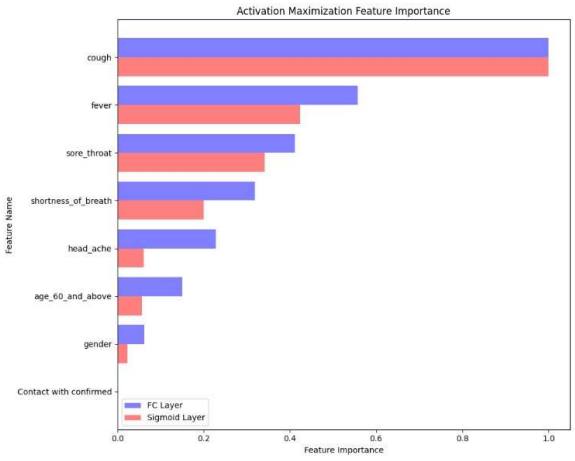
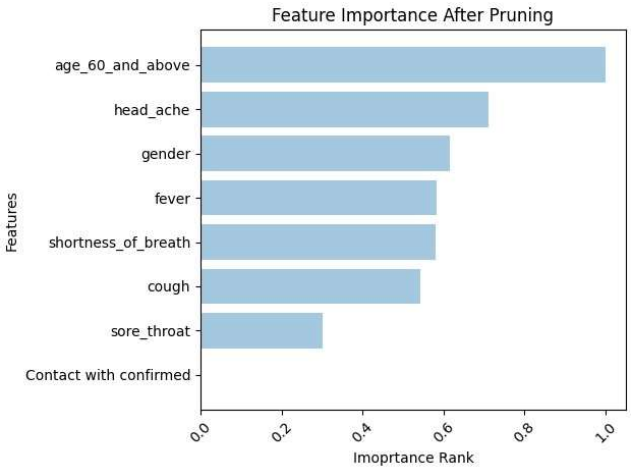


Figure 12: Normalized importance score's covid features Prune Activation.



Where the y-axis refers to the feature importance extraction method and the x-axis shows the normalized importance score as calculated be the normalization' equation . Where is clear that some of the features were giving a normalized score of 1.and the Top features change from one method to another.

III.2 Auc-Roc Curve Covid-19

After obtaining the feature's normalized score's a model that uses only the top 3 features based on the normalized score was produced to check the effect of these features on the model's accuracy thus providing us with measure of evaluation .

Figure 13:Roc curve for each method based on top 3 covid features features

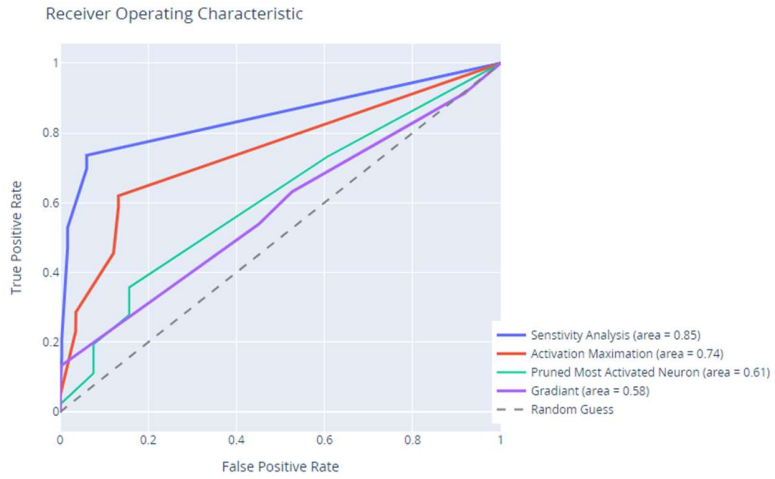


Table 1:Auc scores for each method based on top 3

Feature importance method	AUC Score
Sensitivity	0.85
Activation	0.74
Prune	0.61
Gradient	0.58

Where the Y-axis refer to True positive rate and the X-axis is the False positive rate , while each dataset represents the method off evaluation (**Figure 13**) .The dataset generated with the top 3 features from Sensitivity normalized feature importance's produced the best AUC score as shown in Table 1 (above).

III.3 Blood Stream model

After running bloodstream dataset in the model, a feature importance plot was produced. As can be seen in below where the y-axis refers to the feature importance extraction method and the x-axis shows the normalized importance score as calculated be equation . Where is clear that some methods gave most of features a normalized score of 1 (where 1 is the maximum score meaning this feature has maximum effect on the model forecast). while most gave it for only one feature.

Figure 14: Normalized features Blood Stream model Senstivity

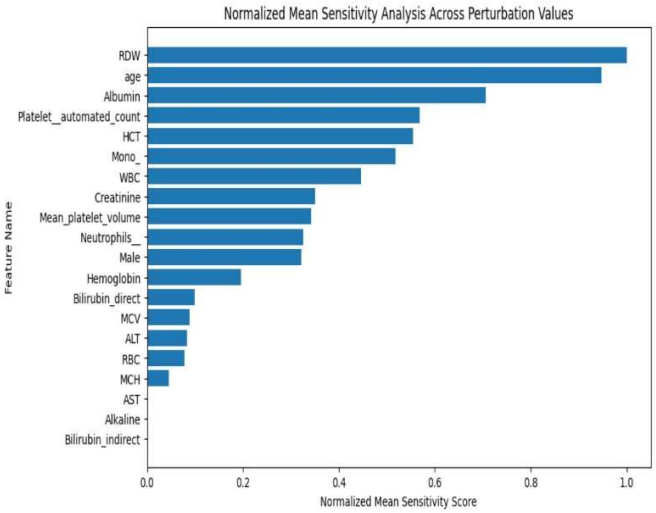


Figure 15: Normalized features Blood Stream model Gradients

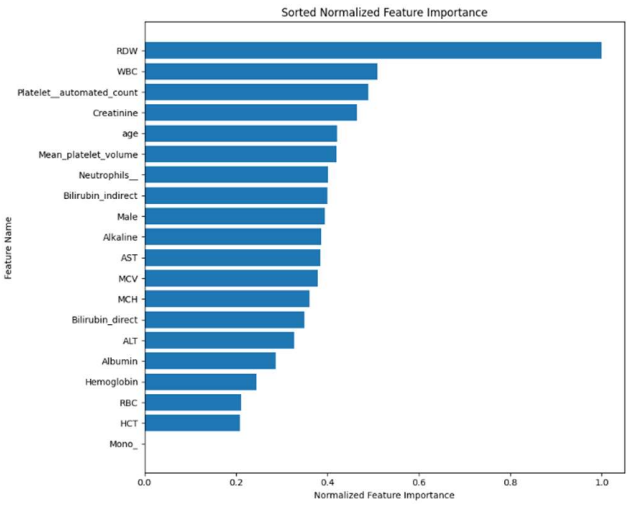


Figure 16: Normalized features Blood Stream model Activation

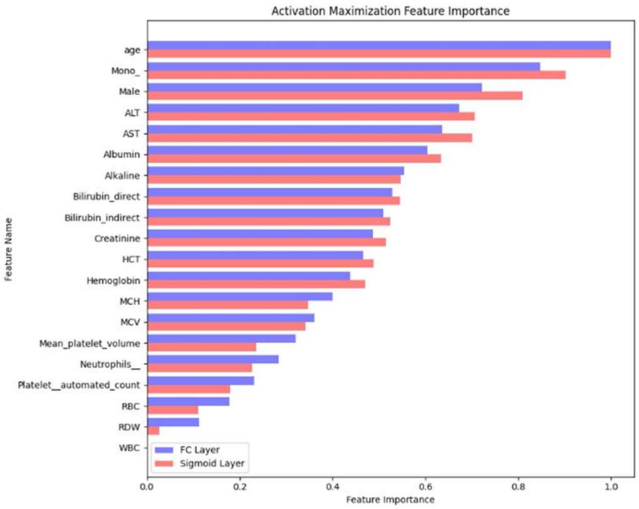
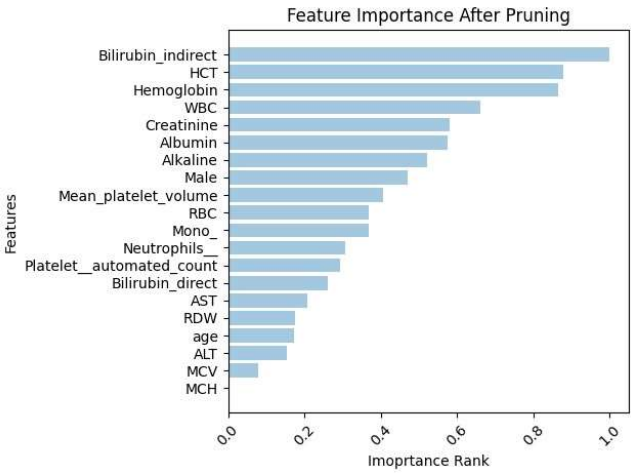


Figure 17: Normalized features Blood Stream model Activation&Pruning



III.4 Auc-Roc Curve Blood Stream

After obtaining the feature's normalized score's a model that uses only the top 5 features based on the normalized score was produced to check the effect of these features on the model's accuracy thus providing us with an evaluation measure**.**

Figure 18: ROC curve for each method based on top 5 bloodstream features

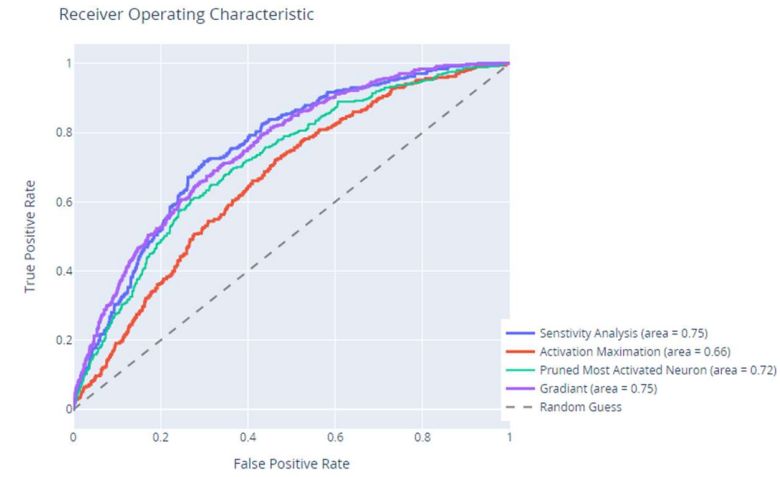


Table 2:Auc scores for each method based on top 5 features

Feature importance method	AUC Score
Sensitivity	0.75
Activation	0.66
Prune	0.72
Gradiant	0.75

Where the Y-axis refer to True positive rate and the X-axis is the False positive rate, while each dataset represents the method off which the top features were selected and the model AUC for those features. And as shown in (Figure 18) both the datasets generated using the top 5 features from Sensitivity and Gradiant normalized feature importance's produced the best AUC score as shown in table 2 (above).

IV. Discussion

As can be seen in figures [*above*Error! Reference source not found.] each method of feature extraction gave a different importance score to each of the features ranging between 0 -1 ,with 1 being the feature with maximum effect on the models forecast**.** To be able to evaluate those methods we rely on the Auc-roc measures mentioned in [Figure 13,Figure 18].The top features selected in covid-19 data set where : fever ,contact with confirmed ,these features offered the best detecting to covid positive result ,and by looking at the PDF of the covid dataset as shown in (Figure 5) we can see that both fever and contact with confirmed both have a higher probability distribution for identifying with a positive corona result which satisfies the pdf assumption. Also, the top features selected in the Bloodstream dataset where: Albumin ,RDW ,Mono and Age. Taking into consideration the PDF of bloodstream shown in (Figure 7) the top features extracted by Sensitivity show a high distribution of being label with high mortality rate (positive for bloodstream infection) meaning that Sensitivity was able to select the top features from each dataset that have the best relation to the prediction and As shown in figures [Figure 9,Figure 14] .In both cases Sensitivity- Permutation analysis produced the highest AUC-ROC curve score thus indicating that the top features selected by thus method provide the best model performance and relation to the prediction as shown in the probability density function.

V. CONCLUSION

Our research provides better insight on the performance of each explanation method when applied to clinical data of different properties. We prove the usefulness of different methods in all these data settings, and the advantage of other methods in some of them. Although those methods are preforming good the development of more causality methods is a necessary step in building more human-like artificial intelligence (AI) for the future, that in turn can learn to deal with real clinical data and produce butter predictions. As AI becomes more advanced, humans are challenged to comprehend and retrace how the algorithm came to a result. The whole calculation process is turned into what is commonly referred to as a “black box" that is impossible to interpret. And that is why a set of processes and methods are being developed like Explainable artificial intelligence (XAI) that allows us to comprehend and trust the results and output created by the DL algorithms. As we move to use more AI models in our decision making the development of methods like XAI is crucial and more research should be done.

Appendix

1. The code used in this project can be found in our GitHub via this link : https://github.com/ShadiKhoury/DL_Project

2. Due to security concerns , and Security agreements singed in order to obtain real word-clinical data , we can't share our own datasets , as this script can be used to implement into your own datasets in seamless manner.

References

[1] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” Knowledge and Information Systems 2013 41:3, vol. 41, no. 3, pp. 647–665, Aug. 2013, doi: 10.1007/S10115-013-0679-X.

[2] G. Schwalbe, B. Finzel, G. Schwalbe, and B. Finzel, “A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts,” *Data Mining and Knowledge Discovery 2023*, pp. 1–59, Jan. 2023, doi: 10.1007/S10618-022-00867-8.

[3] H. Shu and H. Zhu, “Sensitivity Analysis of Deep Neural Networks,” Jan. 2019, doi: 10.1609/aaai.v33i01.33014943.

[4] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “TOWARDS BETTER UNDERSTANDING OF GRADIENT-BASED ATTRIBUTION METHODS FOR DEEP NEURAL NETWORKS.”

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization.” [Online]. Available: http://cnnlocalization.csail.mit.edu

[6] E. Diao, G. Wang, J. Zhang, Y. Yang, J. Ding, and V. Tarokh, “PRUNING DEEP NEURAL NETWORKS FROM A SPARSITY PERSPECTIVE.”

[7] K. Kawaguchi, Z. Deng, X. Ji, and J. Huang, “How Does Information Bottleneck Help Deep Learning?,” May 2023, Accessed: Oct. 07, 2023. [Online]. Available: https://arxiv.org/abs/2305.18887v1
[8] Y. Zoabi, S. Deri-Rozov, and N. Shomron, “Machine learning-based prediction of COVID-19 diagnosis based on symptoms,” *npj Digital Medicine 2021 4:1*, vol. 4, no. 1, pp. 1–5, Jan. 2021, doi: 10.1038/s41746-020-00372-6.
[9] Y. Zoabi, O. Kehat, D. Lahav, A. Weiss-Meilik, A. Adler, and N. Shomron, “Predicting bloodstream infection outcome using machine learning,” *medRxiv*, p. 2021.05.18.21257369, May 2021, doi: 10.1101/2021.05.18.21257369.
[10] E. J. Lannge, “Does removal of correlated variables affect the classification accuracy of machine learning algorithms?”.