iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83585

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Consistency-Aware Evaluation of LIME and SHAP Explanations Using a Hybrid Similarity Metric

Anushka Arora, Hirday Pratap Singh, Mr.Avaneesh Kumar
Department of Computer Science and Engineering Galgotias University, Greater Noida, India

***Abstract*—Explainable Artificial Intelligence methods like LIME and SHAP are used a lot to understand what machine learning models predict.. Most research focuses on generating explanations not checking if they are consistent across different methods. This paper presents a framework called Consistency-Aware Explainable AI that checks how well different explanation meth-ods agree. It uses a score called Explanation Consistency Score, which combines two metrics.**
WetestedthisframeworkontheUCIHeartDiseasedatasetus-ingLogisticRegression,RandomForestandXGBoostclassifiers.
Weevaluatedexplanationconsistencyacross50testinstancesfor each model.
Theresultsshowthatbeinggoodatpredictingdoesnot mean an explanation is consistent. Logistic Regression got the Explanation Consistency Score while Random Forest got the lowest even though it was best at classifying.
TheExplanationConsistencyScoreprovidesanreliablewaytocheckexplanationconsistencyacrossmachinelearningmodels.
TheproposedECSframeworkprovidesancomputationally lightweightapproach,forevaluatingexplanationconsistency
Across machinelearningmodels.
***Index Terms*—Explainable AI, LIME, SHAP, Explanation Consistency, Feature Importance, Jaccard Similarity, Spearman Rank Correlation, Machine Learning**

I. INTRODUCTION

Machinelearningmodelsarebeingusedmoreandmore in areas like healthcare and finance. Some of these modelslike the ones that use different methods together or the ones that are really complex are hard to understand. This makes it toughfor the peopleusingthesemodelstofigureoutwhythey aremakingpredictionswiththemachinelearningmodels.The machine learning models are like a box that you cannot see inside so it is hard to trust the predictions, from the machine learning models. [3].

Explainable Artificial Intelligence is a way to make Ar-tificial Intelligence understandable. It does this by showing howtheArtificialIntelligencemakesitspredictionsbased on the information it gets. There are a couple of ways that people usually do this. One way is called LIME (Local InterpretableModel-AgnosticExplanations)[1].Anotherway is called SHAP (SHapley Additive exPlanations) [2].Even thoughbothLIMEandSHAPtrytoexplainhowtheArtificial Intelligence works they are very different. LIME and SHAP are different because they think about the problem in waysand use different methods to solve it. LIME makes a model just for one prediction at a time. On the hand SHAP uses a concept, from game theory to figure out how much each piece of information contributes to the prediction.

A big question is how often these two methods give the results. We need to know if LIME and SHAP agree with each other.. We have to see if LIME and SHAP agree in different situationslikewithdifferentmodelarchitectures.IfLIMEand SHAPgivedifferentexplanations,forthesamepredictionthen itishardtoknowwhichmethodtousewhenmakingdecisions.

Contributions- This paper makes the following contribu-tions:

- WeintroducetheExplanationConsistencyScore(ECS) which is a hybrid metric combining Jaccard Similarity and Spearman Rank Correlation to measure the overlap between the feature sets and ranking agreement between LIME and SHAP explanations.
- WedesignaCA-XAIevaluationframeworkthatworks with any model and can be used to compare any two methods that explain which features are important. The CA-XAI evaluation framework is really flexible. Can be applied to any pair of feature-importance-based explana-tion methods.
- We are looking at how LIME and SHAP work with LogisticRegression,RandomForestandXGBoostonthe UCIHeart Diseasedataset.Wewanttoseehowconsistent they are when we look at each instance and the whole model.
- Wefound outthat whenweuse complexmodels tomake predictionsitcanbehardertogetconsistentexplanations. This means that there is a balance, between making predictions and getting explanations that make sense for LIME and SHAP.

©IJRASET: All Rights are Reserved

The rest of the paper is organized as follows. Section IIdiscusses the research gap. Section IIIreviews related work. Section IVdescribes the proposed methodology. Section Vdetails the experimental setup. Section VIpresents results and discussion. Section VIIconcludes the paper.

II. RESEARCH GAP AND MOTIVATION

DespitethegrowinguseoftechniqueslikeLIMEandSHAPmostresearchfocusesoncreatingexplanationsnotcheckingifthey'rereliableorconsistentacrossdifferentmethods.Current methodscheckhowgoodexplanationsare,howstrongtheyare or how easy to understand they are but rarely see if different explanation methods give the insights for the same prediction. This is a problem in areas like healthcare and finance where wrongexplanationscanleadtobaddecisions.Moststudiesdo nothaveaway tomeasure ifdifferentexplanationmethods
givesimilarresults.

To fix this we propose a metric called the Explanation Consistency Score (ECS).ECS uses Jaccard Similarity and Spearman Rank Correlation to check if LIME and SHAP explanations are consistent across different machine learning models.

III. RELATED WORK

A. *Explanation Methods*

Ribeiro and other people introduced LIME [1]. LIME is a way to understand how a complex model makes decisions. It does this by using a model that is easier to understand. This simpler model is usually a model with only a few important features. People like LIME because it is easy to use and it works with different types of models.

Lundberg and Lee [2] introduced SHAP. SHAP is a way to figureoutwhichfeaturesareimportant,foramodelsdecisions. It uses ideas from game theory to do this. SHAP is fair and consistentbecauseitfollowssomerules.Theserulesarecalled efficiency, symmetry and additivity. For models that use trees there is an algorithm called TreeSHAP. TreeSHAP makes it possible to calculate these values exactly and quickly [11].

Some people made a list of all the ways to explain how models work. Adadi and Berrada [3] did this. They grouped these methods by what they do how they work with models and what they produce. Guidotti et al. [4] and other peoplealso made a list of ways to explain models. They talked about the challenges of figuring out if these methods are working correctly. They discussed ways to explain models and the problems that come with it.

B. *EvaluationofExplanationQuality*

Evaluatingthequalityofexplanationsisstillsomething we are trying to figure out. We have some ways to measure explanations, like how they match the truth if they are stable and if they are easy to understand.. We do not have one standard way to judge all explanations. The problem is that different methods to explain things can show parts of how a modelworkssoitishardtocomparethemdirectlywithout aclearwaytomeasurethem.Evaluatingthequalityofexplanations is really important. We need a better way to doit.

C. *ConsistencyBetweenExplanationMethods*

The problem of getting explanation methods to agree with eachotherissomethingthatpeoplehavenotlookedinto verymuch.WhenweusemethodslikeLIMEandSHAP on the same model we find that they often pick similar butnot exactly the same important features. However nobody has come up with a way to measure how well these methodsagree with each other. This work is trying to fix this issue by introducing something called ECS, which's a simple and easy to understand way to measure how consistent these methods areandthisthingiscalledaconsistencymetricforexplanation methods, like LIME and SHAP.

IV. PROPOSED METHODOLOGY

Fig. 1shows the process of the proposed CA-XAI frame-work. The CA-XAI framework has four parts.First there is the data preprocessing stage,Then the CA-XAI framework goes through the model training stage and after that it does the explanation generation stage,Finally the CA-XAI framework does the ECS computation stage.

A. *Dataset*

TheUCIHeartDiseaseDataset[12]isusedforexperiments. This dataset has 303 records. It includes 13 features, for each patient.

©IJRASET: All Rights are Reserved

These features are the patients age the patients sex,the type of chest pain the patient has the patients restingbloodpressurethepatientsserumcholesterollevelthepatients fasting blood sugar level the results of the patients resting ECG the patients heart rate whether the patient gets exercise-induced angina the patients ST depression the patients ST slope the number of major vessels the patient has and the patients thalassemia status. The UCI Heart Disease Datasethasatargetvariable.Thisbinarytargetvariableshowswhether the patient has heart disease or not. .

B. ClassificationModels

Threeclassifierswithvaryingcomplexityareevaluated:

- LogisticRegression(LR) [7]: A linear classifier usedas the baseline. Its linear structure makes LIME's local surrogate and SHAP's additive attributions theoretically compatible.
- Random Forest (RF) [5]: A non-linear ensemble of decision trees using bagging. Diversity among trees in-troduces complexity in local explanation approximation.
- XGBoost (XGB) [6]: A gradient-boosted tree ensemble known for high predictive performance. TreeSHAP [11] is used for efficient Shapley value computation.

C. ExplainabilityMethods

1) LIME: LIME generates instance-level explanations by perturbinganinputsample,queryingtheblack-boxmodel on the perturbed samples, and fitting a locally weighted interpretable surrogate.

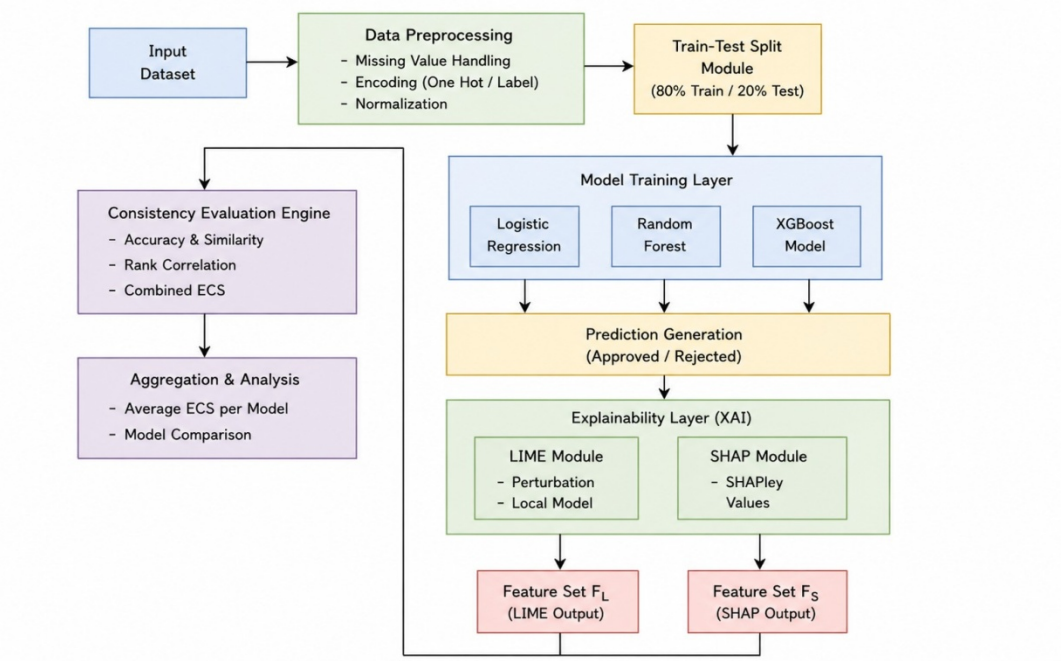


Fig.1.OverallpipelineoftheproposedCA

XAIframeworkincludingpreprocessing,modeltraining,LIME/SHAPexplanationgeneration,andECScomputation.

LIMEinternallygeneratesperturbedsamplesaroundatarget input instance and observes the prediction behavior of the black-box model on these samples. Similarity weights are assigned based on the distance between perturbed samplesand the original instance. A locally interpretable surrogate model is then fitted to approximate model behavior in thelocal neighborhood.

TheoptimizationobjectiveofLIMEis:

$$\operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (1)$$

where $f$is the black-box model, $g$is the surrogate model,$\pi_x$is the locality-aware weighting function, and $\Omega(g)$penal-izes surrogate complexity. The top-k features by coefficient magnitude are extracted from the fitted surrogate.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

2) SHAP: SHAP computes feature importance using Shap-leyvaluesderivedfromcooperativegametheory.Eachfeature contributioniscalculatedastheaveragemarginalcontribution of that feature across all possible subsets of input features.

TheShapleyvalueforfeature$i$is:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (2)$$

where$F$is the complete feature set,$S$is a subsetof features,and$f(S)$is the model output using only featuresin$S$.Fortree-basedmodels,theTreeSHAPalgorithm[11] usedforefficientcomputation.

D. Consistency-Aware XAI Algorithm

TheworkflowoftheproposedCA-XAIframeworkisas follows.

Input:Dataset$D$,machinelearningmodels$M$,top-k feature count k

Output:ECSforeachmodel

1) Loadandpreprocessdataset.
2) Encodecategoricalfeatures;normalizenumericalfeatures.
3) Splitdatasetintotrainingandtestingsets.
4) Trainmodels:LogisticRegression,RandomForest,XG-Boost.
5) Foreachtestinstancex_i:

- GenerateLIMEexplanation;extracttop-$k$featuresF_L.
- GenerateSHAPexplanation;extracttop-$k$featuresF_S.
- ComputeJaccardSimilarity$J$.
- ComputeSpearmanRankCorrelationρ.
- Normalize$\rho$to[0,1].
- Computeinstance-levelECS.

6) AggregateECSscoresacrossallinstances.
7) ReportmeanECSandstandarddeviationpermodel.

E. ExplanationConsistencyScore(ECS)

TheECScapturestwopartsofexplanationagreement.

a) Jaccard Similarity: Let$F_L$be the top-$k$featuresfrom LIME and$F_S$from SHAP. The Jaccard Similarity measures isfeature-setoverlap[8]:

$$J = \frac{|F_L \cap F_S|}{|F_L \cup F_S|} \quad (3)$$

$J \in [0,1]$,where1indicatesidenticalfeaturesets.

b) SpearmanRankCorrelation: Forfeaturessharedbyboth $F_L$and F_S, the Spearman Rank Correlation measures ranking consistency [9]:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (4)$$

©IJRASET: All Rights are Reserved

where d_i is the rank difference for shared feature i and n is the number of shared features. $\rho \in [-1,1]$.

c) ECS Formulation: The Spearman Rank Correlation value is linearly normalized from the range $[-1,1]$ to $[0,1]$ before combining it with Jaccard Similarity:

$$\rho_{norm} = \frac{\rho + 1}{2} \quad (5)$$

The final ECS is computed as:

$$ECS = 0.5 \cdot J + 0.5 \cdot \rho_{norm} \quad (6)$$

ECS $\in [0,1]$, in which the ECS is higher it means that the two explanation methods are more consistent with each other. The ECS gives weight to things, which means it thinks that the identity of the features and the order they are, in are equally important.

F. Computational Complexity

For top-k features, Jaccard Similarity needs $O(k)$ oper-ations. Spearman Rank Correlation requires $O(k \log k)$ op-erations because it has to rank features. The overall ECS computation complexity, per explanation pair is therefore $O(k \log k)$ which makes it lightweight relative to the cost of generating LIME and SHAP explanations.

V. EXPERIMENTAL SETUP

A. Preprocessing

We used one- encoding for the categorical variables. This is a way to make them work with our model. The continuous features were also changed. We made sure they had a mean of zero and a variance of one. This is called z-score normal-ization. We did this before we trained our model. We did it to make sure the continuous features were on the scale, as the categorical variables.

B. Train-Test Split

We took the dataset. Split it into two parts. We used 80 percent for training and 20 percent for testing. This way the class distribution stays the same. When we were picking the settings for the model we used a special method, on the training part. This method is called 5-fold cross-validation.

C. Explanation Generation

LIME and SHAP explanations were made for fifty test instances that were chosen randomly for each model. For each test instance the top ten features that had the biggest impact were picked from each method. This was done by looking at how each feature contributed to the result. The ECS was calculated for each test instance. Then combined together the average and standard deviation for each model. The performance of the classification was checked using measures, like how accurate it was how precise it was how well it recalled things and the F1-score of LIME and SHAP explanations.

D. Implementation Details

The proposed framework was built using Python. I used Google Colab with Python version 3.11. Here are the libraries I used:

- scikit-learn[10] is used for preprocessing and model training
- shap is used for feature attribution generation
- lime is used for local explanation generation
- xgboost is used for gradient-boosted classification
- scipy is used for Spearman rank correlation computa-tion

TA used a fixed seed of 42 everywhere to make sure the results can be repeated. The following hyperparameters were used:

- LogisticRegression:max_iter=1000
- RandomForest:default scikit-learn parameters
- XGBoost:eval_metric=logloss

VI. RESULTS AND DISCUSSION

A. *Classification Performance and ECS*

Table Ishows how well the models did with classification. ItalsogivesustheECSvaluesforeachofthethreemodels.

TABLEI CLASSIFICATIONPERFORMANCEANDEXPLANATIONCONSISTENCYACROSSMODELS

Model	Acc.	Prec.	Rec.	F1	MeanECs	StdECS
LogisticReg.	0.8533	0.9000	0.8411	0.8696	0.5576	0.0930
RandomForest	0.8913	0.9223	0.8879	0.9048	0.3879	0.1744
XGBoost	0.8750	0.9038	0.8785	0.8910	0.5125	0.1289

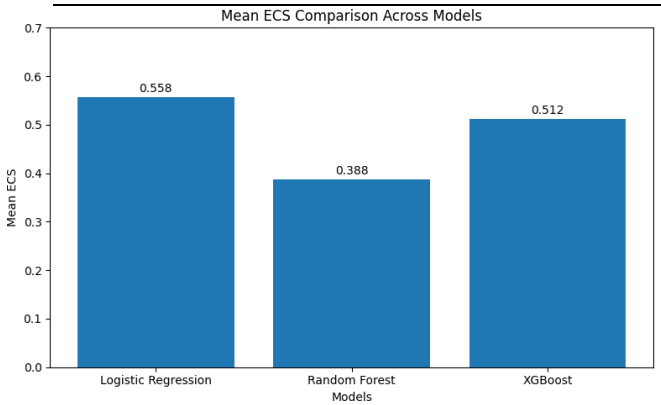


Fig. 2.Mean ECS comparison across Logistic Regression, Random Forest,and XGBoost models. Logistic Regression achieved the highest explanationconsistency, while Random Forest exhibited the lowest ECS despite higherpredictive accuracy.

B. *Analysis*

The results of the experiment show that the explanations given by machine learning models are not consistent. Logistic Regression gave the consistent explanations with a score of 0.5576. This means that the explanations from LIME and SHAP mostly agreed with each other for Logistic Regression. XGBoost had a score of 0.5125 which's pretty good.. Ran-dom Forest had the lowest score of 0.3879 which means its explanations were not very consistent.

WhatisinterestingisthatRandomForestwasreallygoodat making predictions with an accuracy of 89.13 percent.. It had the lowest explanation consistency score. This shows that just because a model is good at making predictions it does not mean that its explanations will always make sense. Models like Random Forest that combine different models can makeit hard for LIME and SHAP to agree on which features are important.Thisisbecausethesemodelshaverules,formaking decisions.

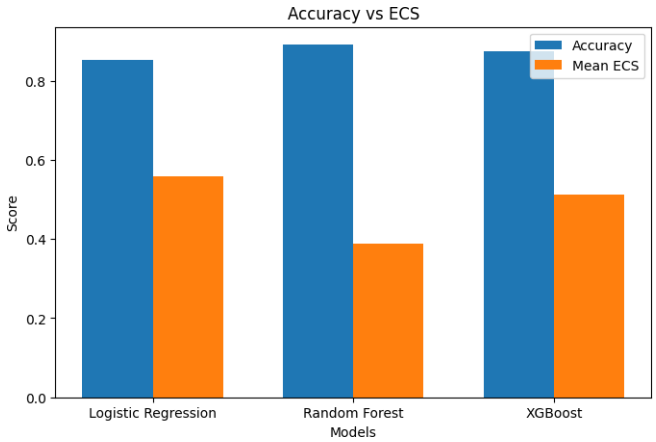


Fig.3.ComparisonbetweenclassificationaccuracyandmeanECSacrossLogistic Regression, Random Forest, and XGBoost models

In contrast Logistic Regression gave predictions but had consistent explanations. This is because it uses an clear method.
Both LIME and SHAP should give results for simple modelslikeLogisticRegression.LIMEmakesasimplemodel to mimic the one and SHAP adds up the effects of each feature.Theseresultsshowthatyoucan'thavebothpredictions and clear explanations at the same time. This is important when choosing a model, for applications where understanding the model matters.

C. ExplanationStability

The scores from ECS give us an idea of how stable the explanationsre when we test them. Logistic Regression hadthe amount of variation in ECS scores, which is 0.0930. This meansthattheexplanationsfromLogisticRegressionaremore stable when we look at samples. On the hand Random Forest hadthebiggestamountofvariationinECSscores,which is0.1744.

ThissuggeststhattheexplanationsfromRandom Forest are not as consistent and can change a lot when we use models that combine many things. XGBoost had an amountof variation, in ECS scores, which is 0.1289.
The heart disease dataset is something that the classifiersre really good at predicting.If you look at the confusion matrices in Fig. 4you can see that all three classifiers do a job. The RandomForestheartdiseaseclassifieristhebest,atgettingthe positive instances right.The Logistic Regression heart disease classifier is also good because it makes predictions and does not make a lot of unstable classifications. The heart disease dataset is predicted well by all three classifiers.

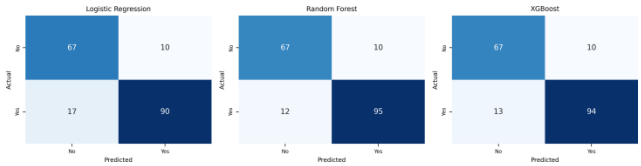


Fig.4.ConfusionmatricesforLogisticRegression,RandomForest,andXGBoost classifiers evaluated on the UCI Heart Disease dataset.

Fig. 5shows how ECS values are spread out across the test instances we sampled. Logistic Regression has a distribution that is more grouped with less variation which means that LIMEandSHAPexplanationsagreewitheachotherinaway. Random Forest shows wider dispersion which suggests less stable explanation consistency across all samples.

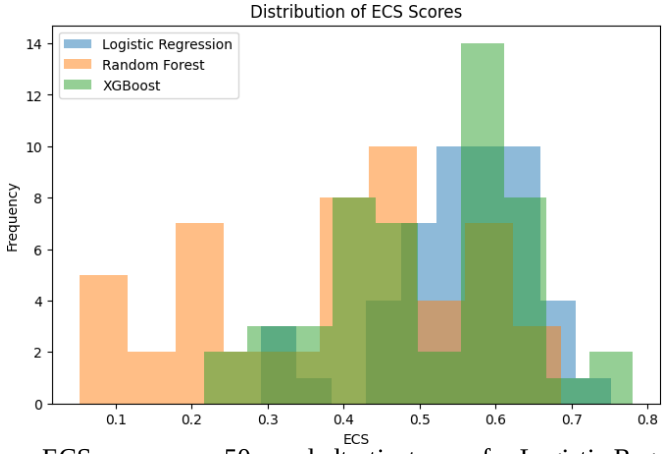


Fig.5.Distributionofper-instanceECSscoresacross50sampledtestinstances for Logistic Regression, Random Forest, and XGBoost models.

Fig. 6shows how much the explanations from LIME and SHAP agree on each feature. If we look at things like the type of chest pain the ST slope and if someone gets angina when they exercise we can see that LIME and SHAP explanations usually agree on these things. This means that these features areimportantforthemodelsandtheyareimportanteverytime. The features like chest pain type and ST slope are consistent. This is true, for LIME explanations and SHAP explanations.

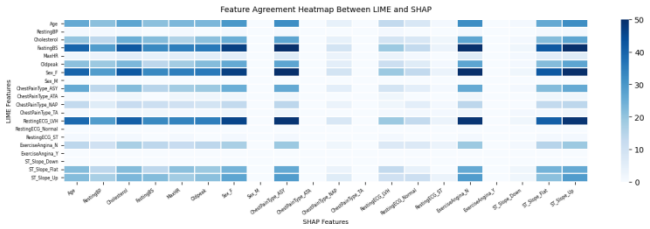


Fig.6.

VII. CONCLUSION AND FUTURE WORK

This paper is about the Explanation Consistency Score, which's a way to measure how well two explanation methods, LIME and SHAP agree with each other. The Explanation Consistency Score is a mix of two things: Jaccard Similarity and Spearman Rank Correlation.

TheresearcherstriedouttheExplanationConsistencyScore with a few models, like Logistic Regression, Random Forest andXGBoostonadatasetaboutheartdisease.Theyfoundthat Logistic Regression had the Explanation Consistency Score, which was 0.5576 and it was also very stable. On the hand Random Forest was really good at predicting things but its Explanation Consistency Score was the lowest, at 0.3879.

Thismakesusthinkthatmaybemodelsthataretoocomplex do not do a job of explaining things in a consistent way. So when we are choosing a model we should not just think about how it predicts things but also about how well it explains things.

The good thing about the Explanation Consistency Score is that it is not hard to compute and it works with any model. Itis also easy to add to the way we already evaluate models.

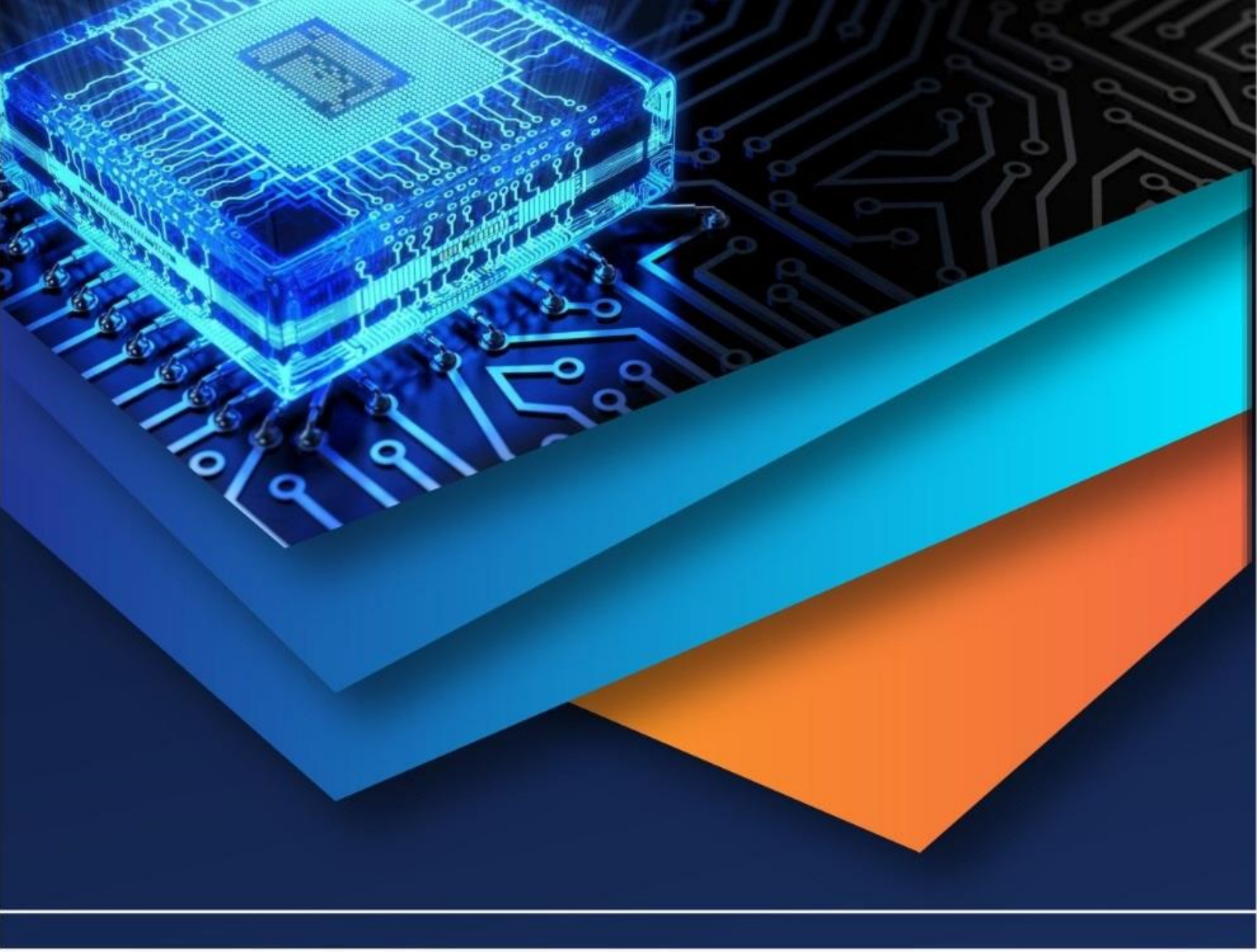
There are a things to keep in mind though. The researchers only tried this out on one dataset so we do not know if it will workthewayonotherdatasets.TheyalsoonlyusedLIMEand SHAPsowedonotknowwhatwouldhappenwithexplanation methods.. They gave equal weight to Jaccard Similarity and Spearman Rank Correlation which might not always be the best thing to do.

Theresearchersdideverythingtheycouldtomakesuretheir results are reliable, by using the random seeds every time and making their code public.

In the future the researchers want to try the Explanation Consistency Score with complex models like deep neural networks and see if they can make it work better by changing the way they weigh Jaccard Similarity and Spearman Rank Correlation.

REFERENCES

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?:Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDDInt.Conf.KnowledgeDiscoveryandDataMining,pp.1135–1144,2016.
[2] S. M. Lundberg and S.-I. Lee, "A Unified Approach to InterpretingModel Predictions," in Advances in Neural Information ProcessingSystems (NIPS), vol. 30, pp. 4765–4774, 2017.
[3] A.AdadiandM.Berrada,"PeekingInsidetheBlack-Box:ASurveyon Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp.52138–52160, 2018.
[4] R.Guidotti,A.Monreale,S.Ruggieri,F.Turini,F.Giannotti,and Pedreschi,"ASurveyofMethodsforExplainingBlackBoxModels,"ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2019.
[5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp.5–32, 2001.
[6] T.ChenandC.Guestrin,"XGBoost:AScalableTreeBoostingSystem,"in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and DataMining, pp. 785–794, 2016.
[7] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed.New York, NY: Wiley, 2000.
[8] P. Jaccard, "The Distribution of Flora in the Alpine Zone," NewPhytologist, vol. 11, no. 2, pp. 37–50, 1912.
[9] C.Spearman,"TheProofandMeasurementofAssociationBetweenTwoThings," American Journal of Psychology, vol. 15, no. 1, pp. 72–101,1904.
[10] F.Pedregosaetal.,"Scikit-learn:MachineLearninginPython,"Journalof Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[11] S. M. Lundberg et al., "From Local Explanations to Global Understand-ing with Explainable AI for Trees," Nature Machine Intelligence, vol.2, no. 1, pp. 56–67, 2020.
[12] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart DiseaseDataset,"UCIMachineLearningRepository, 1988.[Online]. Available:https://archive.ics.uci.edu/ml/datasets/heart+disease



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)