Bulletin of Electrical Engineering and Informatics
Vol. 11, No. 2, April 2022, pp. 1126~1133
ISSN: 2302-9285, DOI: 10.11591/eei.v11i2.3391

Extraction of human understandable insight from machine learning model for diabetes prediction

Tsehay Admassu Assegie[1], Thulasi Karpagam[2], Radha Mothukuri[3], Ravulapalli Lakshmi Tulasi[4], Minychil Fentahun Engidaye[1]
[1]Department of Computer Science, College of Natural & Computational Science, Injibara University, Injibara, Ethiopia
[2]Department of AI & DS, R.M.K College of Engineering & Technology, Kavara ipettai, Tamil Nadu, India
[3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India
[4]Department of Computer Science and Engineering, R.V.R & J.C College of Engineering, Guntur, India

Article Info

Article history:

Received Nov 20, 2021
Revised Feb 24, 2022
Accepted Mar 5, 2022

Keywords:

Diabetes prediction
Extreme boosting (XGBoost)
Interpretable model
LIME
SHAP

ABSTRACT

Explaining the reason for model’s output as diabetes positive or negative is crucial for diabetes diagnosis. Because, reasoning the predictive outcome of model helps to understand why the model predicted an instance into diabetes positive or negative class. In recent years, highest predictive accuracy and promising result is achieved with simple linear model to complex deep neural network. However, the use of complex model such as ensemble and deep learning have trade-off between accuracy and interpretability. In response to the problem of interpretability, different approaches have been proposed to explain the predictive outcome of complex model. However, the relationship between the proposed approaches and the preferred approach for diabetes prediction is not clear. To address this problem, the authors aimed to implement and compare existing model interpretation approaches, local interpretable model agnostic explanation (LIME), shapely additive explanation (SHAP) and permutation feature importance by employing extreme boosting (XGBoost). Experiment is conducted on diabetes dataset with the aim of investigating the most influencing feature on model output. Overall, experimental result evidently appears to reveal that blood glucose has the highest impact on model prediction outcome.

Corresponding Author:

Tsehay Admassu
Department of Computer Science, College of Natural & Computational Science, Injibara University
P.O.B: 40, Injibara, Ethiopia
Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

In prediction of diabetes with ensemble learning model, the claim that highest possible accuracy is achieved by employing complex ensemble learning does not make sense to domain expert or nephrologist [1]. This is because the predictive outcome of ensemble model or the claim that a patient is suffering from diabetes based on the predictive outcome of model do not have sense to domain expert in terms of predictive accuracy. Hence, there is a need for building model that is reliable and interpretable. In response to model interpretability problem different approaches such as local interpretable model explanation (LIME), Shapely additive explanation (SHAP) and permutation based global model interpretation approaches such as random forest and extreme boosting feature importance are widely employed to interpret the predicted outcome of ensemble model in medical domain [2]. As the amount of data produced by healthcare centers is growing rapidly, the use of predictive analytics and predictive model for patient diagnosis requires the use of complex models to discover insight from large datasets through knowledge discovery process [3]. However, the predictive outcome of ensemble model is not easily interpretable despite their higher accuracy.

Interpretation and explanation of ensemble model for diabetes identification is the process where we understand the predictive outcome of the model. In model interpretation, we try to explain why the model predicted an instance in diabetes dataset as diabetes positive or negative and why the accuracy is lower or higher. Hence, model explanation or interpretation deals with reasoning why a particular instance is being classified as diabetes positive or negative locally or the global feature influence to the model's predictive outcome. In addition to that, model explanation is one of the method, which is gaining much focus by machine learning researchers in recent years. Because, the gap between model explanation or interpretability and detailed understanding of intelligent application is making difficult to distinguish between domain expert's knowledge and the data presented to ensemble model, predicted outcome and accuracy achieved by the model.

The objective of this research is to discuss the state of the art and future research directions of interpretable machine learning. In addition, the authors aimed to implement the state of art interpretability approaches and explain how different features of diabetes dataset influence the prediction whether or not patient has diabetes, based on different diabetes diagnostic measurements such as blood glucose level, body mass index, age, insulin, diabetes pedigree function (DPF), blood pressure, skin thickness and number of times (pregnancies) a woman is pregnant. Overall, this research investigates the risk factor that has high influence on getting diabetes. Thus, this study investigates the answers to the following research questions: i) which diabetes feature has highest influence on the predicted outcome of extreme boosting model?; ii) what is the model explanation method that provides better insights to extreme boosting model output?; iii) How to bridge the gap between explanations of individual model predictions to explanation truly characterizing the general model behavior?

The rest of this study is organized as: in section 2, the state of the art model interpretation approaches are discussed. In section 3, the research methodology and the dataset employed for implementing and testing the existing interpretability approach is presented. Section 4 presents the experimental results of the research and discusses the result obtained, at the end section 5 concludes the research.

2. RELATED WORK

In recent years, ensemble based predictive model excels the human expert in diabetes prediction. Machine learning and artificial intelligent model have become indispensable tool for tasks such as disease prediction. Today's machine learning models are achieving excellent performance even exceeding the human intelligence [4]. However, complex machine learning models are not providing an information about how the model has arrived at particular prediction outcome. In the medical domain, lack of transparency is not acceptable; the development of methods for feature importance visualization, explaining and interpreting complex model has recently attracted the attentions of many researchers [5]. Thus, this research presents some of the recent methods and applications in the field of machine learning model explanation and implementation of interpretable model and to apply the existing methods of explainable learning algorithms using the diabetes dataset.

Model explanation is crucial to reason model's predicted outcome such as positive or negative class in the medical domain [6], [7]. To reason and get insight into the dataset feature, SHAP and local interpretable model explanation is the most common method employed for interpreting the predicted outcome of machine learning model. In [8], local interpretable model explanation and SHAP is compared for model interpretation and comparative result reveals that SHAP is better for estimating the influence of feature on predicted outcome of machine learning model as compared to LIME.

In another study [9], the authors developed interpretable extreme boosting model with SHAP explanation. The authors conducted experiment on the performance of the developed model and the model shows area under curve of 0.71 for diabetes prediction. While the developed model is found to be interpretable, the performance is lower and have much scope for improvement. Moreover, body mass index (BMI) is the most influencing feature to the model's predicted outcome. In contrast, diabetes pedigree function (DPF) or family history has lowest influence on the predicted outcome of extreme gradient boosting model. As future work, the authors suggested the work to be extended using ensemble learning model and the work does not have comparison to other state of the art model explanation approach such as local interpretable model agnostic explanation (LIME).

In another work [10], SHAP based feature importance is applied to classification problem for model interpretation. Although, machine-learning model is achieving acceptable performance, the interpretation and explainability of predictive outcome of machine learning model is ongoing research. Moreover, using human expert knowledge for building intelligent model makes predictive outcome reliable.

In [11], [12], comparative performance analysis is conducted on diabetes prediction using decision tree and random forest model. In the study, the authors applied model based or embedded feature selection with random forest and decision tree model. The experimental result evidently reveals that random forest with feature selection achieved good performance as compared to decision tree model. Overall, the highest accuracy achieved by random forest model is 90%. However, the authors does not provided interpretability of their model and the reason behind why the model performed with 90% accuracy is not clear for domain expert.

The use of explainable method in predictive model development is crucial to reason the predicted outcome of tree based model. In addition, model explanation is important to identify the risk factor of diabetes disease. In [13], [14], the authors applied SHAP and developed an explainable extreme boosting model for diabetes prediction. The experimental result reveals that good diabetes prediction performance is obtained with extreme boosting model. The experimental result shows that area under curve (AUC=71.13) score is achieved using the extreme boosting model.

Literature survey on machine learning model used for diabetes prediction shows that artificial neural network is applied to diabetes disease prediction in [15]. As shown in the experimental result of the study artificial neural network model achieved accuracy of 93%. However, the study does not employed model explanation method for reasoning the predicted outcome of artificial neural network model. While the model achieved better accuracy, the study lacks interpretability for reasoning each and overall predictive outcome of the model.

Several supervised diabetes detection model exists for diabetes diagnosis in the literature [16], [17]. Most recently, the researchers are investigating the topic of explanation of predictive model and method applied for model explanation and interpretation. Even though higher accuracy is achieved on diabetes prediction using automated intelligent model, explanation remains an issue and method that is more sophisticated is required to explain why particular model achieved such higher accuracy.

3. METHOD

To conduct this study, the authors collected diabetes dataset consisting of 768 observations openly available to the scientific community from Kaggle data repository. The dataset consists of 268 samples of diabetes positive and 500 samples of diabetes negative. In the implementation and experimental test, XGBoost algorithm is employed. Thus, automated diabetes prediction model is developed by using XGBoost which is explained by using Shapley additive (SHAP) and LIME.

To classify diabetes dataset, the authors developed XGBoost model. However, the developed model should provide reasonable outcome. Hence, to explain the outcome of the model, Shapley Additive explanation (SHAP) and LIME is employed. With SHAP and LIME the contribution of each feature to model output is explained telling why the model arrived at diabetes positive or negative outcome. The additive influence of each input feature on XGBoost model is defined by the formula given in (1) [18]-[20].

$$f(x1, x2\ \ldots xm) = \sum_{1=1}^{\mathrm{m}} \alpha_{\mathrm{i}}) \qquad (1)$$

Where (x1, x2 …xM) is the input feature and αi denotes the SHAP value for the impact of feature i for specific or local estimate y= $f(x1, x2\ \ldots xm) = \sum_{1=1}^{\mathrm{m}} \alpha_{\mathrm{i}})$. Here, i indicates an instance and the predicted outcome of each input feature is calculated and the weighted sum is used to explain the influence of each feature to model output.

4. RESULTS AND DISCUSSION

This section presents the experimental results obtained by using three different model explanation method such as SHAP, LIME and XGBoost permutation based feature importance. SHAP is employed to explain local and global influence of diabetes feature whereas LIME and XGBoost feature importance provided the global feature importance on model output.

4.1. SHAP for feature influence on XGBoost model output

The experiment is conducted to analyze the SHAP values to interpret the impact of having a certain value of diabetes feature for a given feature in comparison to the prediction outcome taking the base line value for the diabetes feature value. To get an insight into which diabetes features are most important for developed extreme boosting model predictive outcome, the SHAP plot, which shows values of every diabetes feature for every sample, is demonstrated in Figure 1. The summary plot shown in Figure 1 reveals which diabetes features are most important, and also their range of effects over the diabetes dataset.

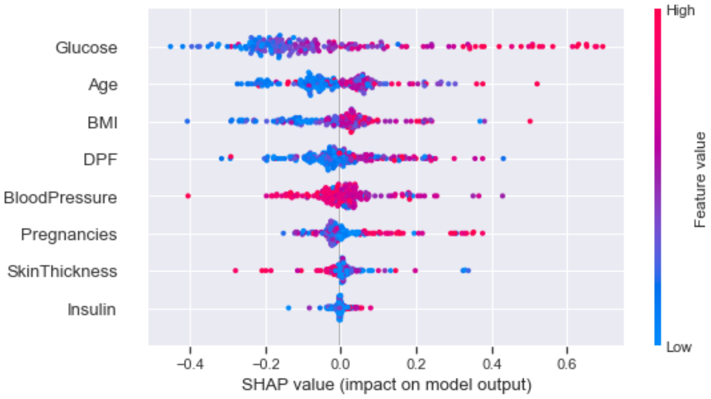


Figure 1. Explanation of XGBoost with SHAP

As demonstrated in Figure 1, the vertical location shows the diabetes feature the SHAP explanation is representing; the color shows whether that feature was high or low for that row of the diabetes dataset sample. The horizontal location shows whether the effect of that value caused a higher or lower prediction. The point in the upper left shows a person whose glucose level is less reduces the prediction of diabetes by the extreme boosting model by 0.4. In contrast, a person with highest glucose level shows an increase by 0.6 by the extreme boosting model predicted as diabetes positive. The result confirms from Figure 1, that high values of glucose play positive important role in the final prediction, and diabetes values of insulin seems a contrary to have a negative impact on predicting diabetes positive. In addition to the global explanation shown in Figure 1. The local feature importance or influencing on a given instance prediction by the XGBoost model is demonstrated in Figure 2.

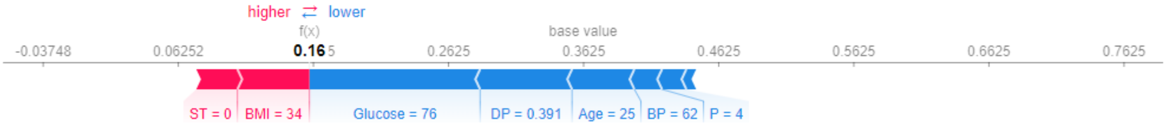


Figure 2. Local explanation using SHAP

As ilustrated in Figure 3, the most influencing feature for extreme boosting model output is glucose level. This means, patient with higher glucose level has highest probaility to being classified as diabetes positive class. Age is the second most influencing feature affecting the mangnitude of the exterme boosting model output. In contrast, inuslin has lowest influence on model output or predidictive ouctome. Skin thickness is the second least influence model output as we observe from Figure 3.

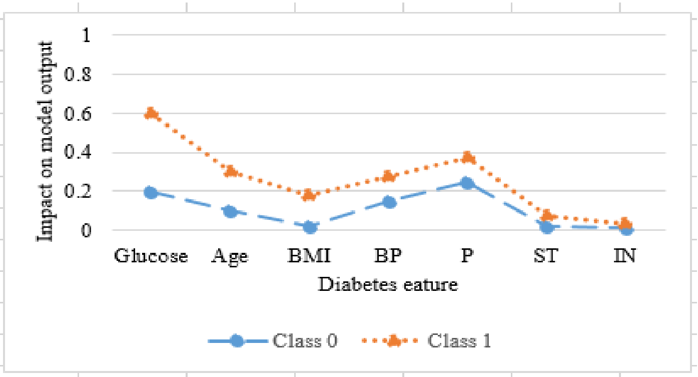


Figure 3. Mean SHAP value average impact of feature on model output magnitude

4.2. XGBoost for feature influence explanation

The authors conducted an experiment on the global diabetes feature importance calculations that come with XGBoost based on the following parameters: i) feature weights: based on the number of times a feature appears in a tree across the ensemble of trees; ii) coverage: the average coverage (number of samples affected) of splits which use the feature; iii) gain: the average gain of splits, which use the feature. The global diabetes feature importance based on feature weights, coverage and gain is demonstrated in Figure 4.

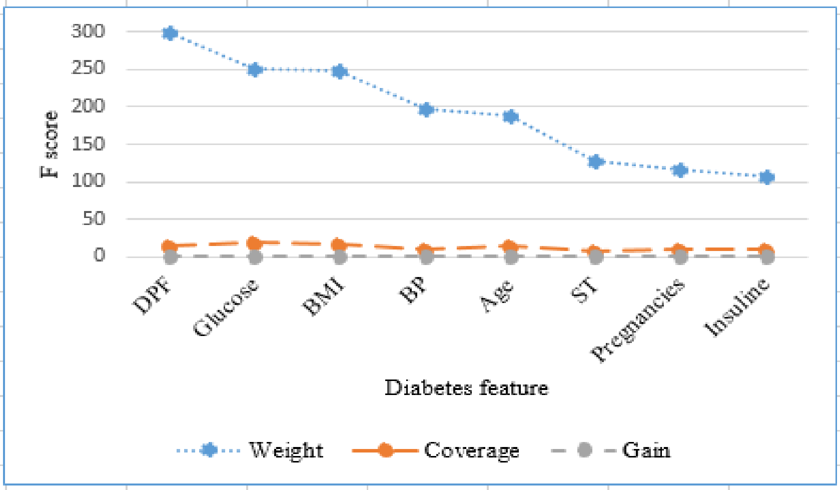
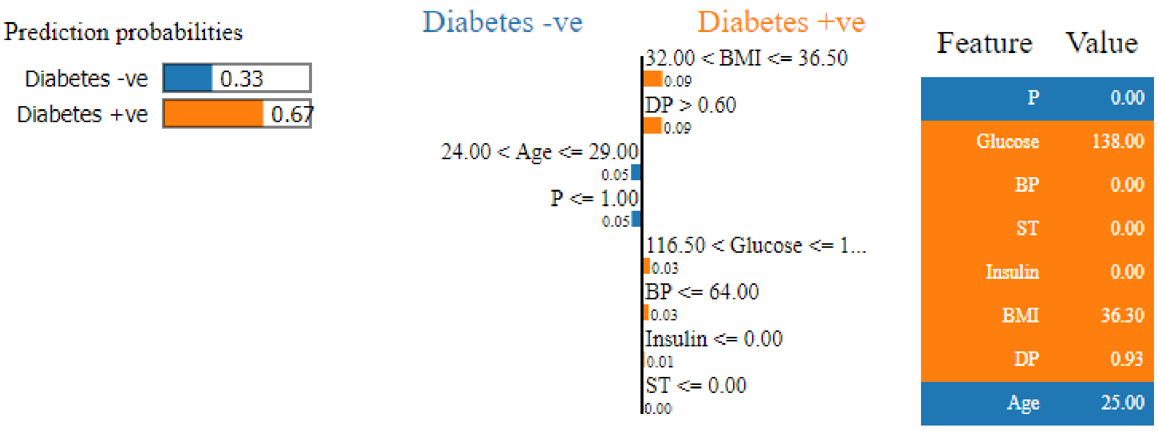


Figure 4. XGBoost permutation based feature importance

As demonstrated in Figure 4, diabetes pedigree function is the most important feature based on the weight. Based on feature coverage, glucose is the most important feature as compared to other diabetes features used in the dataset. Similarly, based on feature split mean gain value glucose is the most important feature that plays positive role in diabetes prediction.

4.3. LIME for feature influence on XGBoost model output

In addition to global explanation provided by XGBoost feature importance, LIME is employed to explain local influence of diabetes features on extreme boosting model output. Figure 5 illustrates the explanation generated using LIME to explain the local positive prediction outcome of XGBoost model. As demonstrated in Figure 5, glucose, blood pressure (BP), skin tackiness (ST), diabetes pedigree function (DPF), body mass index (BMI) and insulin level has positive influence on the model prediction output as diabetes positive. In contrast, pregnancies and age have negative impact on positive predicted output.



Feature	Value
P	0.00
Glucose	138.00
BP	0.00
ST	0.00
Insulin	0.00
BMI	36.30
DP	0.93
Age	25.00

Figure 5. Explanation generated by LIME

4.4 Receiver operating characteristic curve of XGBoost

Area under curve (AUC) is the area enclosed under the receiver operating characteristic curve (ROC) [21]-[25]. The AUC score is widely employed to evaluate binary classifier. AUC score is interpreted as follows: AUC score=0.5 shows random classifier. Whereas AUC score of 1.0 is a perfect classifier and AUC score of less than 0.5 is poor quality. The AUC score of our model is demonstrated in Figure 6. As shown in Figure 4, the AUC score of the developed model is 0.82 which is promising result.

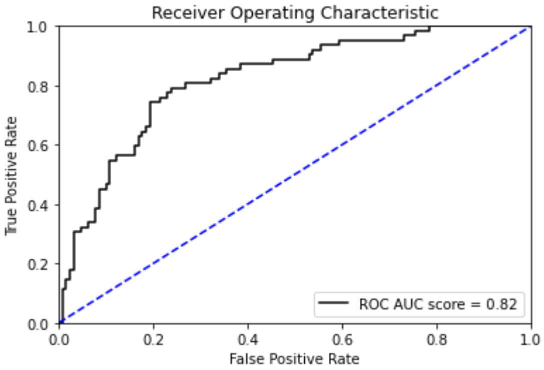


Figure 6. AUC score of XGBoost

5. CONCLUSION

In this research, more reliable model is developed for diabetes prediction by employing extreme boosting (XGBoost) algorithm with SHAP. The predictive outcome the model is interpreted with SHAP, LIME and permutation based feature importance. The experimental result appears to prove that model interpretation is significantly important to develop more accurate diabetes prediction model. In addition, model explanation provides an information whether a given feature contributes to prediction outcome of model or not. Overall, with the existing model explanation method such as SHAP, LIME and permutation based feature selection, XGBoost is interpretable and convincing to domain experts why the model has made the prediction it has made on a given instance. SHAP provides local and global effect of diabetes features while permutation based feature importance with XGBoost and LIME only provide the global influence of feature to model output.

In the feature work, we plan to work with casual effect of the diabetes feature. The model interpretation method discussed in this study does not provide counterfactual effect of a feature to the model’s predicted outcome. Thus, the approaches that provide an information on what measures could be taken to reduce the effect of diabetes by employing causal inference is recommended for future work.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the internet facilities and laptop provided by Injibara University to conduct experiment and complete this work, at Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia.

REFERENCES

[1] T. A. Assegie and P. S. Nair, “The Performance of Different Machine Learning Models on Diabetes Prediction,” *International Journal of Scientific & Technology Research*, vol. 9, no. 1, pp. 2491-2494, January 2020.

[2] L. Lama *et al.*, “Machine learning for prediction of diabetes risk in middle-aged Swedish people,” *Heliyon*, pp. 1-6, June 2021, doi: 10.1016/j.heliyon.2021.e07419.

[3] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou and K. S. Nikita, "An explainable XGBoost–based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus," 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), 2020, pp. 859-864, doi: *10.1109/BIBE50027.2020.00146*.

[4] Y. Xiang *et al.*, “Artificial Intelligence-Based Diagnosis of Diabetes Mellitus: Combining Fundus Photography with Traditional Chinese Medicine Diagnostic Methodology,” *Hindawi BioMed Research Internationa*l, vol. 2021, p. 5556057, 2001, doi: 10.1155/2021/5556057.

[5] K. K. Chari, M. C. Babu, and S. Kodati, “Classification of Diabetes using Random Forest with Feature Selection Algorithm,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 1, 2019, doi: 10.35940/ijitee.L3595.119119.

[6] S. J. Sushma, T. A. Assegie, D. C. Vinutha and S. Padmashree, “An improved feature selection approach for chronic heart disease detection,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3501-3506, 2021, doi: 10.11591/eei.v10i6.3001.

[7] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, “An Improved Artificial Neural Network

Model for Effective Diabetes Prediction," *Hindawi Complexity*, vol. 2021, p. 5525271, 2021, doi: 10.1155/2021/5525271.

[8] R. Patil and S. Tamane, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes*," International Journal of Electrical and Computer Engineering*, vol. 8, no. 5, October 2018, pp. 3966-3975, doi: 10.11591/ijece.v8i5.pp3966-3975.

[9] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning based model for predicting hypertension," *BMC Medical Informatics and Decision Making*, vol. 19, no. 146, 2019, doi: 10.1186/s12911-019-0874-0.

[10] E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. A. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," *arXiv:2002.11097*, 2020.

[11] C. S. Kumar, M. N. S. Choudary, V. B. Bommineni, G. Tarun and T. Anjali, "Dimensionality Reduction based on SHAP Analysis: A Simple and Trustworthy Approach," *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 558-560, doi: 10.1109/ICCSP48568.2020.9182109.

[12] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021, doi: 10.3390/ electronics10050593.

[13] A. Saadallah and K. Morik, "Active Sampling for Learning Interpretable Surrogate Machine Learning Models," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 264-272, doi: 10.1109/DSAA49011.2020.00039.

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

[15] M. A. Ahmad, A. Teredesai and C. Eckert, "Interpretable Machine Learning in Healthcare," *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, pp. 447-447, doi: 10.1109/ICHI.2018.00095.

[16] J. -X. Mi, A. -D. Li and L. -F. Zhou, "Review Study of Interpretation Methods for Future Interpretable Machine Learning," in *IEEE Access*, vol. 8, pp. 191969-191985, 2020, doi: 10.1109/ACCESS.2020.3032756.

[17] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, March 2021, pp. 184-190, doi: 10.11591/ijai.v10.i1.pp184-190 184.

[18] T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," *Journal of Robotics and Control*, vol. 2, no. 3, May 2020, doi: 10.18196/jrc.2363.

[19] N. B. Kumarakulasinghe, T. Blomberg, J. Liu, A. Saraiva Leao and P. Papapetrou, "Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models," *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 7-12, doi: 10.1109/CBMS49503.2020.00009.

[20] I. C. Covert, S. Lundberg, and Su-In Lee, "Explaining by Removing: A Unified Framework for Model Explanation," *Journal of Machine Learning Research*, vol. 22, pp. 1-90, 2021.

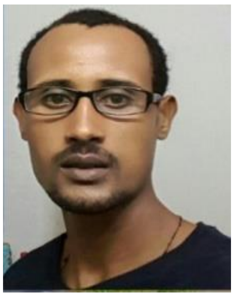
[22] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models", *SN Applied Science*, vol. 3, no. 272, 2021, doi: 10.1007/s42452-021-04148-9.

[23] Batunacun, R. Wieland, T. Lakes, and C. Nendel, "Using SHAP to interpret XGBoost predictions of grassland 2 degradation in Xilingol, China," *Geo Scientific Model Developmen*t, pp. 1-28, 2020, doi: 10.5194/gmd-2020-59.

[24] R. Rodriguez-Perez and J. Bajorath," Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," *Journal of Computer-Aided Molecular Design*, vol. 34, pp. 1013–1026, 2020, doi: 10.1007/s10822-020-00314-0.

[25] S. N. Payrovnaziri *et al*., "Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review," *Journal of the American Medical Informatics Association*, vol. 27, no. 7, 2020, pp. 1173–1185, doi: 10.1093/jamia/ocaa053.

BIOGRAPHIES OF AUTHORS



Tsehay Admassu Assegie received the B.Sc., degree in Computer Science from Dilla University, Dilla Ethiopia in 2013. He received Master degree in Computer Sceince from Faculty of Science Andhra Univeristy, India in 2016. Currently He is working as Lecturer in the Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia. His research interests include machine learning, medical image analysis and pattern recongition. He has published 28 research articles in international reputed and peer reviewed journals. He can be contacted at email: tsehayadmassu2006@gmail.com

Thulasi Karpagam is currently working as assistant professor in the Department of Artificial Intelligence and Data Science at R.M.K College of Engineering and Technology, Kavaraipettai, Chennai, India. Her Research areas include cloud computing, machine learning and big data analytics. She can be contacted at email: karpagamdv83@gmail.com

Radha Mothukuri is currently working as an Assistant professor in the department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India. She submitted her thesis to Acharya Nagarjuna University. Her area of research interest are Text mining, machine learning, cloud computing and network security. She can be contacted at email: radhahemanth12@gmail.com.

Ravulapalli Lakshmi Tulasi is currently working as a Professor in the Department of Computer Science and Engineering, R.V.R & J.C College of Engineering, Guntur, Andhra Pradesh, India. Her research interests include machine learning, data mining, information retrieval systems, and semantic Web. She can be contacted at email: rtulasi.2002@gmail.com

Minychil Fentahun Engidaye is currently working as lecturer in the Department of Computer Science, Injibara University, Injibara, Ethiopia. His research interest include machine learning and natural language processing. He can be contacted at email: minychil@gmail.com