

Improving breast cancer prediction through explainable artificial intelligence – A transdisciplinary approach

Reena Lokare¹, Jyoti Sunil More², Vaishali V. Sarbhukan³, Mansing Rathod¹,
Sarita Rathod¹, Sunita Patil⁴

¹Department of Information Technology, K.J. Somaiya Institute of Technology, Mumbai, India

²Department of Computer Engineering, Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India

³Department of Information Technology, Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India

⁴SVKM's NMIMS, Shirpur Campus at SVKM's Narsee Monjee Institute of Management Studies (NMIMS), Dhule, India

Article Info

Article history:

Received Aug 24, 2024

Revised Mar 17, 2025

Accepted Jul 3, 2025

Keywords:

Artificial intelligence
Breast cancer
Explainable AI
healthcare
prediction

ABSTRACT

Artificial intelligence (AI) technology has shown tremendous contributions in various applications like speech recognition, expert systems, computer vision, robotics, and gaming. machine learning (ML) and deep learning (DL) algorithms under AI address problems such as prediction, classification, and regression. AI has touched many domains. The results or the predictions generated by these algorithms are not easily accepted by the user. Especially, the Healthcare domain is facing a great challenge in accepting the results or the predictions with the concern, Are AI results reliable, correct, and ethical? Doctors or medical practitioners are not ready to treat patients based on results or suggestions generated by AI algorithms. Hence, a technology that can explain how the results returned by AI algorithms are trustworthy, transparent, and interpretable was strongly needed. This need has given rise to the latest technology-explainable artificial intelligence (XAI). With the use of XAI, all the predictions, classifications made by AI algorithms are explainable, auditable, comprehensive, validating, and socially acceptable. This paper discusses explaining the results of breast cancer prediction as a case study. The results show that such an explanation will build trust in the doctors and hence will increase the acceptance of the AI-based systems.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Reena Lokare

Department of Information Technology, K.J. Somaiya Institute of Technology
Mumbai, India

Email: reena.l@somaiya.edu

1. INTRODUCTION

Around 85% of organizations in 2020 used artificial intelligence (AI) and this number is consistently growing [1]. These recent developments in the field of AI are notable. It has not left a single field untouched. The world is behind automation. Every sector wants to automate their businesses and processes. Banking, Manufacturing Industries, Automobile, Healthcare to name a few. The reasons for this automation can be to reduce manpower, increase efficiency in their process, fast outcome, future predictions, or suggestions, solving complex problems like disease prediction, object detection, and fraud detection. Every sector looking for automation through AI technology is introducing intelligence in their traditional applications.

Many complex problems in the healthcare sector such as disease prediction and detection, drug discovery, and precision medicine recommendations. are being solved by the AI algorithms [2]. For example, an AI application to predict cancer disease in a patient may give prediction in terms of an accuracy.

This prediction can be made based on attributes or features such as age, symptoms, tumor stage, and number of tumors observed. Predictions in AI/machine learning (ML)/deep learning (DL) are nothing but algorithms/models' "guesses," about the target value based on given attributes/features. These modern complex algorithms are naturally opaque, non- intuitive and difficult to understand. From this opaqueness, one might guess but can never know what is happening inside the network. The models are trained and tested with huge data samples, getting a good accuracy with a convincing positive predictive value. Calculating only accuracy to explain the real-life tasks is not sufficient.

Medical professionals such as doctors/clinicians/medicinal experts are not ready to accept results generated by AI algorithms with the biggest concern, 'Predictions made by AI are trustworthy?' How to trust the results generated by ML algorithms? Applying an unexplained network with patients' life is near to impossible. In order to trust AI generated results, a doctor needs an explanation of the result. They want to understand the way AI works to help them improve their services as no one really knows why these algorithms make the choices they do? This lack of transparency is the biggest challenge for the healthcare sector in using AI predicted results.

Most of the AI algorithms are complex in nature. They generate predictions or classifications based on the features provided. One never knows what is happening inside the network. Cioffi *et al.* [3], trusting predictions made by AI algorithm based on confusion matrix is not at all acceptable by Healthcare, Finance, Department of Defense (DoD) like automated target recognition, battlefield surveillance, autonomous vehicles, smart homes, and smart buildings. as well as nearly by every domain. They need justification to trust the results, because it is a great concern for people's lives, health, safety, security, and dependability.

ML algorithms like support vector machine (SVM), linear regression (LR), neural network (NN), decision tree (DT), and random forest (RF) algorithm. generate results in the form of prediction accuracy, True/False or Yes/No. Black box is a term used for the model where the working of the model is hidden and not understood by the user [4]. The model decisions are non-transparent and often incomprehensible even to the experts or the computer developers as they do not explain the results. Thus, the interpretability, reliability, resiliency, trustworthiness and explainability is expected by all complex applications using AI/ML/DL model predictions.

Liao *et al.* [5] found that the major limitation of the AI algorithm is their black box nature. In order to accept the results generated by these algorithms, the technology called explainable artificial intelligence (XAI) came into existence. This technology has the capability to explain the results generated by AI models. Though XAI is vastly under development, it has strength to bridge the gap between researchers and machines. There are many challenges in using interpretable ML technology as indicated by Rudin *et al.* [6]. It includes optimization of DTs, optimization of scoring systems, and placing constraints into models for better interpretability. XAI technology will play a major role in the Healthcare sector where there is still a little acceptance of AI-predicted results. Complex diseases such as cancer, diabetes, heart diseases, and many other hereditary diseases are common worldwide. Changing lifestyle, self-medication, wrong eating habits are some of the reasons behind these diseases. Though many treatment options are available and being practiced in many countries, accurate and most suitable treatment, and timely disease predictions are some of the challenges [7]. Use of AI technology is drastically helping in predicting the future, generating expected results, providing suggestions in such very complex problems. The invention of XAI, along with the powerful implications of AI works well, making AI technology serve humankind to a great extent.

XAI is a recent research field where the solutions provided by AI algorithms are explained. With XAI, the solutions provided by a NN are examined and possibly interpreted by an expert. Interpretable/explainable/transparent ML is a set of tools and frameworks or methods for understanding the predictions of AI algorithms [8]. Explainable AI and transparent AI are the terms used interchangeably. XAI is a technology that makes the ML models interpretable. If the AI model itself is interpretable, all predictions made by the model is themselves explainable. XAI converts black box AI model into a white box model. It introduces transparency in the solutions provided. However, it is still vastly under development to bridge the gap between researchers and machines. Interpretability/explainability can be defined as,

- Knowing the reason behind the decision.
- Maintaining consistency in model's prediction.
- Making predictions based on the feature values.
- Understand model's outputs for classification and regression tasks.
- Shows contribution of each feature in terms of feature importance.

Higher the interpretability, it is easier to understand the decisions or predictions made by the model. Few of the benefits achieved through explanations are increased social acceptance, understanding the science behind the solutions generated by AI algorithms, increased research purpose, accuracy, fidelity, consistency, stability, comprehensibility, certainty, degree of importance, novelty, and representativeness. Users can verify whether the model is behaving as expected, recognize biases in models, and get ideas for

ways to improve the model and training data. These explanations are also useful to the developer. Bertossi and Geerts [9] states that developers and data scientists can verify whether the model is behaving as expected, recognize biases in the models and get ideas for ways to improve the model and training data. Interpretability make AI model predictions transparent and thus increases acceptance of the results by building the trust of users. As models become more complex, the task of producing an interpretable version of the model becomes more difficult. So, it is true that one explanation does not fit all [10].

Intrinsic model-specific methods inspect model components as a path in DT or a rule in decision rule or a weight of a feature. From the regression models, consider the LR model, it predicts the target as a weighted sum of the feature inputs where features could be numeric, binary, categorical or intercept type. While logistic regression model, interprets classification probabilities with two possible outcomes. Instead of plotting straight line, a logistic function is used to squeeze the output of a linear equation between 0 and 1. The DT finds the relationship between features and outcome which are nonlinear. Interaction among features is considered. The decision rule is a simple IF-THEN statement with condition and then prediction. Few more intrinsic model specific methods are RuleFit, generalized linear model (GLM), generalized additive model (GAM), Naïve Bayes, and K-nearest neighbour (KNN). as mentioned by Lin and Chang [11]. Guidotti *et al.* [12] stated few tools or methods which are designed for model agnostic interpretations are: reversed time attention model (RETAIN) which works on recurrent neural networks (RNN), layer-wise relevance propagation (LRP), local interpretable model-agnostic explanations (LIME), partial dependence plot (PDP), individual conditional expectation/impact confidence ease (ICE), permutation feature importance, accumulated local effects (ALE) plot, feature interaction, scoped rules (Anchors), ELI5, Kernel SHAP.

2. BREAST CANCER PREDICTION USING GENOMIC DATASET- A CASE STUDY

Breast cancer is the common type of cancer seen in India and world-wide [13]. Cancer prognosis is possible using Gene expression dataset. Rigorous studies are being carried out across the world to find cancer biomarkers. Cancer disease needs patient specific treatment [14]. Common treatment does not suit every patient. Das *et al.* [15] explains that gene expression is the perfect biomarker for giving patient specific treatment. Researchers have found specific genes that mutate in specific types of cancers. If a patient performs gene testing for specific genes, cancer can be detected at an early stage. Detection of cancer at an early stage increases the chances of a patient's survival. Pathologies like Metropolis in India are offering gene testing at an affordable cost.

2.1. Dataset source

To explain the AI/ML model explainability and reliability in the healthcare domain, we have taken Breast Cancer gene expression dataset from UCI ML repository of ML databases (Centre for ML and intelligent systems) from National Institute of Diabetes and Digestive and Kidney Diseases [16]. The objective is diagnostically predicting whether a patient has breast cancer malignancy or not, based on certain diagnostic measurements included in the dataset. From the breast cancer gene expression dataset, 21 genes are i.e., BRCA1, BRCA2, PALB2, CHEK2, CDH1, PTEN, STK11, TP53, ATM, BARD1, BRIP1, CASP8, CTLA4, CYP19A1, FGFR2, LSP1, MAP3K1, MRE11A, NBN, RAD51, TERT are taken. These genes are tested using Microarray testing as stated by [17]. This dataset contains data of 1,078 patients.

2.2. Prediction tests

Different ML algorithms are tested on this dataset. SVM is giving 97.68 % prediction accuracy with 99% precision, 96% recall and 97% F1-score. Naïve Bayes algorithm gives 92.59% prediction accuracy with 94% precision, 91% recall and 92% F1-score. DT algorithm gives 98.98% prediction accuracy with 96% precision, 95% recall and 96% F1-score. K-NN algorithm gives 93.05% prediction accuracy with 100% precision, 86% recall and 92% F1-score.

The performance of ML algorithms mentioned above shows that they all are performing well. Any software developer can easily trust these algorithms for generic applications. But the doctors, clinicians or medical experts who treat patients, conduct healthcare research, develop medicines and vaccines are based on their knowledge, experience and mainly patient's historical data with disease symptoms, age, and many other factors. Disease prediction for a patient using AI/ML technology must readily be understood by medical practitioners of all experience levels. They should understand the cause of decisions, be able to relate feature values of instance to its prediction, know the reason behind classification and prediction results, and to understand which feature has contributed in the prediction and what value of feature was considered in the prediction. ML algorithms must explain to the doctors why a patient is predicted cancer positive or negative bringing trust in decisions. The machine predicted results should be auditable, comply, validate, and debug.

3. INTERPRETATION OF MACHINE LEARNING MODELS WITH EXPLAINABLE AI

In this section we will see how the predictions or results or suggestions given by AI/ML models are explained or interpreted for complex healthcare applications using the recent technology explainable AI. The transparency and the trustworthiness of intrinsic model-specific and post-hoc i.e., model-agnostic models as mentioned by Mitros and Mac Namee [18] are presented with the case study on breast cancer dataset. For an intrinsic model-specific method, a simple LR model is explained which itself is explainable with the help of available functionalities. Such explanations are not possible with all the ML models due to their complex nature. For predicting results of post-hoc model-agnostic explanations, LIME is interpreted.

3.1. Intrinsic model-specific linear regression model

LR model gives the output from the sum of the weights of the feature inputs. Kim *et al.* [19] mentions linear models are easiest to interpret. Model-specific explanations are easily possible with the LR algorithm. LR can be formulated as shown in (1).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \epsilon_i \quad (1)$$

Where Y_i is the target, X_i represents the input feature, β_0 is called the intercept term and is the constant. β_i represents the weight of the input on the final output. ϵ is called the noise of the model. The task here is to train, test and explain that a patient's breast cancer is benign or malignant with the given dataset, applying LR with Python. Initially relevant libraries and packages are loaded, the dataset is read and viewed, columns and observations are seed from the dataset. Using the stats model package from Python coefficients, t-statistics, standard errors, and p-values for each feature are measured as shown in the Figure 1. The confidence intervals values of features are identified as shown in Figure 2. The coefficient's values and standard errors can be plotted as shown in Figure 3.

```
model.summary()
```

| OLS Regression Results | | | | |
|----------------------------|------------------|---------------------|---------|-------|
| Dep. Variable: | CANCER | R-squared: | 0.795 | |
| Model: | OLS | Adj. R-squared: | 0.791 | |
| Method: | Least Squares | F-statistic: | 194.5 | |
| Date: | Mon, 25 Jan 2021 | Prob (F-statistic): | 0.00 | |
| Time: | 10:48:22 | Log-Likelihood: | -70.899 | |
| No. Observations: | 1078 | AIC: | -97.80 | |
| Df Residuals: | 1056 | BIC: | 11.82 | |
| Df Model: | 21 | | | |
| Covariance Type: nonrobust | | | | |
| | coef | std err | t | P> t |
| const | 0.9297 | 0.047 | 19.772 | 0.000 |
| BRCA1 | 0.0059 | 0.015 | 0.400 | 0.690 |
| BRCA2 | 0.0054 | 0.011 | 0.474 | 0.635 |
| PALB2 | 0.1714 | 0.019 | 8.811 | 0.000 |
| CHEK2 | -0.1160 | 0.016 | -7.074 | 0.000 |
| CDH1 | 0.0167 | 0.004 | 4.052 | 0.000 |
| PTEN | -0.0790 | 0.014 | -5.730 | 0.000 |
| STK11 | 0.2500 | 0.017 | 14.473 | 0.000 |
| TP53 | -0.0368 | 0.012 | -3.147 | 0.002 |
| ATM | -0.0821 | 0.017 | -4.809 | 0.000 |
| BARD1 | 0.0460 | 0.015 | 3.042 | 0.002 |
| BRIP1 | 0.0108 | 0.010 | 1.071 | 0.284 |
| CASP8 | -0.0450 | 0.018 | -2.478 | 0.013 |
| CTLA4 | 0.0415 | 0.012 | 3.434 | 0.001 |
| CYP19A1 | 0.0029 | 0.009 | 0.324 | 0.746 |
| FGFR2 | -0.0246 | 0.008 | -3.190 | 0.001 |
| LSP1 | 0.0690 | 0.015 | 4.681 | 0.000 |
| MAP3K1 | 0.0122 | 0.009 | 1.342 | 0.180 |
| MRE11A | -0.0482 | 0.029 | -1.650 | 0.099 |
| NBN | 0.0863 | 0.016 | 5.517 | 0.000 |
| RANBP1 | 0.1896 | 0.013 | 14.518 | 0.000 |

Figure 1. Summary of LR's output using stats model package

```
model.conf_int()
```

| | 0 | 1 |
|-------|-----------|-----------|
| const | 0.837424 | 1.021954 |
| BRCA1 | -0.022988 | 0.034742 |
| BRCA2 | -0.016793 | 0.027498 |
| PALB2 | 0.133209 | 0.209538 |
| CHEK2 | -0.148173 | -0.083822 |
| CDH1 | 0.008634 | 0.024847 |
| PTEN | -0.106006 | -0.051921 |
| STK11 | 0.216068 | 0.283843 |
| TP53 | -0.059725 | -0.013853 |
| ATM | -0.115619 | -0.048613 |
| BARD1 | 0.016330 | 0.075705 |

Figure 2. Confidence interval values of features

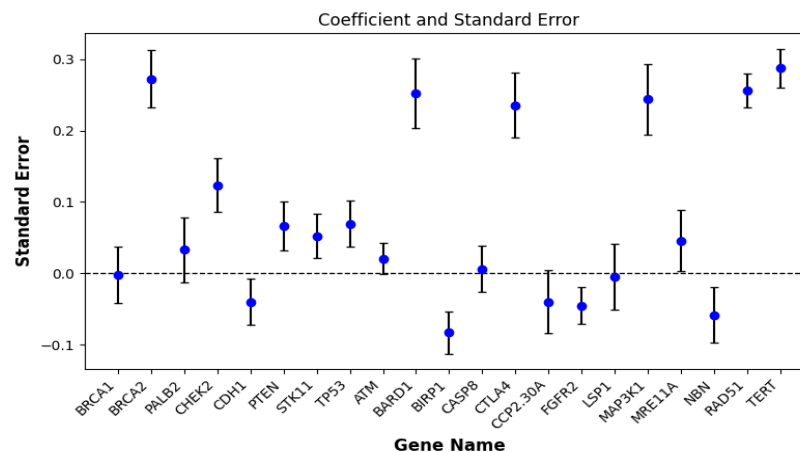


Figure 3. Coefficient values and standard error plot

3.2. Post-hoc model-agnostic LIME model

Model-agnostic methods are more flexible and more usable compared to model-specific methods. These methods can be applied to any ML model. LIME is an example of a post-hoc model-agnostic method. LIME can explain the predictions made by any ML model like LR, and DT classifier. It provides an individual and local explanation of a decision after it has been made. It modifies a single data sample by tweaking the feature values and observes the resulting impact on the output by creating a new dataset having samples permuted and the new predictions on the model.

The interpretable model is trained on this newly generated dataset. This new model has less loss as it finds the closeness of the explanation to the prediction of the initial model. Thus, LIME gives explanations with the help of the contribution of each feature for the prediction and finds which feature changes will have the impact on the prediction [20], [21]. Here let see how the predictions made by the DT classifier are explained with the help of the LIME model as shown in Figure 4. The genes that are used for making the prediction by the classifier are shown along with the gene's values. Also, the genes that positively contribute in the predictions are shown in green color and one which negatively contribute in the prediction are shown in orange color. The LIME model has a notebook feature which explains prediction results in more detail as shown in Figure 5.

This is the prediction made for one instance. The notebook shows features that positively contribute to the predictions and it also shows features that negatively contribute to the predictions. It also shows which features are used for the prediction and what values of the features are used. It helps the doctor to understand and find which genes are mutated. If specific genes are mutated in a patient, it indicates that the patient is likely to develop cancer soon. In case of higher chances of developing cancer, treatment can be started immediately with the help of more testing, precautions, and medications [22]-[25]. By generating the classification report and predicting the macro average, weighted average accuracy with precision, recall and F measure the suggested results become explainable.

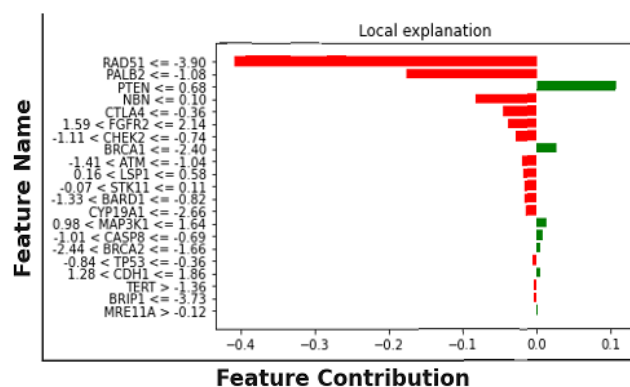


Figure 4. Local explanations of DT classifier by LIME

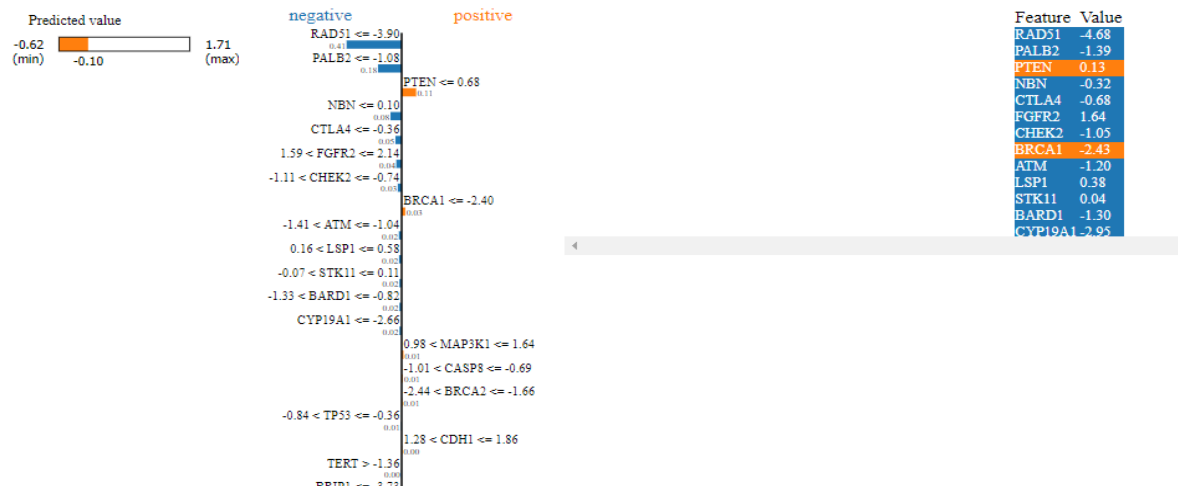


Figure 5. LIME Notebook explaining predictions made by the DT classifier

4. RESULTS

The implemented results for intrinsic model-specific and post-hoc model agnostic explainable AI methods conclude that AI/ML algorithm predictions can be made usable, trustworthy, reliable, and more importantly acceptable in today's complex applications like healthcare 4.0. XAI models show that the results of any ML algorithm can be explained to all levels of users, making them trust and believe the correctness of predicted results. XAI packages are available in Python as well as R programming languages changing the black-box nature of ML algorithms into completely transparent white-box models.

5. CONCLUSION

Today every business sector, industry or organization is implementing Industry 4.0 standards with demand for auditing, trust, and reliability. The explainable/Interpretable transparent/reliable AI presented in this chapter focusing healthcare 4.0 is a step towards building confidence in AI technological trends for the betterment of humankind. Transparency also leads to acceptance of AI technology especially in healthcare. This chapter has presented the scope for adopting AI in complex applications like healthcare with a concern about harm to oneself and to a society. Chapter focuses on the expectations from AI results to support reasoning and evidence for predicted outcomes and finally provides the solution in terms of explainable AI serving this job. Any human being is always curious towards the world, i.e., towards knowing the reasons behind every event. Hence XAI is such technology which helps the user know the facts, hidden knowledge in the AI model's predictions. Once these facts are known to the user, trust is built and acceptance of AI also increases. Also, it is clear from the explanation whether the AI model is making decisions ethically. Hence, in high-impact domain such as healthcare use of XAI have become mandatory. But there are some situations where explanations are not important such as problems where solutions will not have much impact: take example of AI based chess game played for fun. In this case, any decision taken by the AI game is not going to create any impact on the user. Hence, the predictions won't have any impact like loss or gain. So domains such as entertainment, games, fun are termed as low-impact causing areas and doesn't demand any explanations for AI model's predictions. Whereas domains such as healthcare, finance are high-impact, causing areas that demand for explanations.

Also, when problems are studied well e.g., Character recognition application. It is not expected in such an application to explain the predictions made. The problem is really well-designed and doesn't require any explanation. Either the character is classified correctly or wrongly. If AI model is not able to classify the character correctly, model tuning need to be done in order to improve the prediction accuracy and when models are not expected to manipulate. There are models which are not expected to manipulate and have just assistance, informative, analytical roles such as descriptive models, diagnostic models, predictive models, and recommendation models. Thus, one can understand the importance of XAI in high-impact domain. Use of XAI ensures explainability which helps in building trust and increase acceptance of the AI models. It also ensures ethical standards as everything is now interpretable.

One of the disadvantages of XAI technology lies in its interpretable nature. Take an example of a Credit card fraud detection system. If the explanations of such a system will be understood by the fraud person only, then he will take care not to take the identified steps and will find out other ways of doing fraud.

This alternate fraud might not get detected by the underlying model. Thus, with such numerous cases a research scope is open for aspirants who want to see a better concrete AI technological future with explainable AI.

ACKNOWLEDGEMENTS

This research is supported by Dr. Vivek Sunnapwar, Principal, K. J. Somaiya Institute of Technology, Mumbai. We thank him for all the guidance and support provided in this research. We also thank our family members for their kind support.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|--------------------|---|---|----|----|----|---|---|---|---|---|----|----|---|----|
| Reena Lokare | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | ✓ | |
| Jyoti More | | ✓ | | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Vaishali Sarbhukan | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | ✓ | ✓ |
| Mansing Rathod | ✓ | ✓ | | | | ✓ | ✓ | | | | | | | |
| Sarita Rathod | | | | | ✓ | | ✓ | | | ✓ | | ✓ | | ✓ |
| Sunita Patil | ✓ | ✓ | | | | ✓ | | | | ✓ | | | | ✓ |

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

DATA AVAILABILITY

The data used in the research is publicly available to access at <https://archive.ics.uci.edu/>.




REFERENCES

- [1] S. Spreeuwenberg, "Choose for AI and for explainability," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11878 LNCS, 2020, pp. 3–8.
- [2] J. Borana, "Applications of artificial intelligence & associated technologies," *Proceeding of International Conference on Emerging Technologies in Engineering, Biomedical, Management and Science [ETEBMS-2016]*, pp. 64–67, 2016.
- [3] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice, "Artificial intelligence and machine learning applications in smart production: progress, trends, and directions," *Sustainability*, vol. 12, no. 2, p. 492, Jan. 2020, doi: 10.3390/su12020492.
- [4] R. Arun, "Explainable AI: from black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137–141, 2020.
- [5] Q. V. Liao, M. Singh, Y. Zhang, and R. Bellamy, "Introduction to explainable AI," in *Conference on Human Factors in Computing Systems - Proceedings*, May 2021, pp. 1–3, doi: 10.1145/3411763.3445016.
- [6] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, no. none, pp. 1–85, Jan. 2022, doi: 10.1214/21-SS133.
- [7] J. Zugazagoitia, C. Guedes, S. Ponce, I. Ferrer, S. Molina-Pinelo, and L. Paz-Ares, "Current challenges in cancer treatment," *Clinical Therapeutics*, vol. 38, no. 7, pp. 1551–1566, Jul. 2016, doi: 10.1016/j.clinthera.2016.03.026.
- [8] F. K. Dosilovic, M. Bric, and N. Hlupic, "Explainable artificial intelligence: a survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215, doi: 10.23919/MIPRO.2018.8400040.
- [9] L. Bertossi and F. Geerts, "Data quality and explainable AI," *Journal of Data and Information Quality*, vol. 12, no. 2, pp. 1–9, Jun. 2020, doi: 10.1145/3386687.




- [10] R. Hamon, H. Junklewitz, and I. Sanchez, "Robustness and explainability of artificial intelligence," *Joint Research Centre (European Commission)*, p. 40, 2020, [Online]. Available: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC119336/dpad_report.pdf?ref=https://githubhelp.com.
- [11] Y. Lin and X. Chang, "Towards interpreting ML-based automated malware detection models: a survey," *ArXiv preprint*, 2021, [Online]. Available: <http://arxiv.org/abs/2101.06232>.
- [12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.
- [13] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: a survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/electronics8080832.
- [14] S. Ganesh N, D. Rahul, S. Jyotsana, S. Piush, and S. K.K., "Various Types and management of breast cancer: an overview," *Journal of Advanced Pharmaceutical Technology and Research*, vol. 1, no. 2, pp. 109–126, 2010.
- [15] J. Das, K. M. Gayvert, and H. Yu, "Predicting cancer prognosis using functional genomics data sets," *Cancer Informatics*, vol. 13s5, p. CIN.S14064, Jan. 2014, doi: 10.4137/CIN.S14064.
- [16] M. Lichman, "UCI machine learning repository," University of California, School of Information and Computer Science, 2016.
- [17] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Science Signaling*, vol. 6, no. 269, Apr. 2013, doi: 10.1126/scisignal.2004088.
- [18] J. Mitros and B. Mac Namee, "A categorisation of Post-hoc explanations for predictive models," *ArXiv preprint*, Apr. 2019, [Online]. Available: <http://arxiv.org/abs/1904.02495>.
- [19] H. Kim, W.-Y. Loh, Y.-S. Shih, and P. Chaudhuri, "Visualizable and interpretable regression models with good prediction power," *IIE Transactions*, vol. 39, no. 6, pp. 565–579, Mar. 2007, doi: 10.1080/07408170600897502.
- [20] M. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016, pp. 97–101, doi: 10.18653/v1/N16-3020.
- [21] M. R. Zafar and N. M. Khan, "DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *ArXiv preprint*, 2019, [Online]. Available: <http://arxiv.org/abs/1906.10263>.
- [22] F. Gabbay, S. Bar-Lev, O. Montano, and N. Hadad, "A LIME-based explainable machine learning model for predicting the severity level of COVID-19 diagnosed patients," *Applied Sciences*, vol. 11, no. 21, p. 10417, Nov. 2021, doi: 10.3390/app112110417.
- [23] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: a comparative study of local machine learning interpretability techniques," *Computational Intelligence*, vol. 37, no. 4, pp. 1633–1650, Nov. 2021, doi: 10.1111/coin.12410.
- [24] S. Dey *et al.*, "Human-centered explainability for life sciences, healthcare, and medical informatics," *Patterns*, vol. 3, no. 5, p. 100493, May 2022, doi: 10.1016/j.patter.2022.100493.
- [25] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, Jan. 2022, doi: 10.3390/diagnostics12020237.

BIOGRAPHIES OF AUTHORS






Reena Lokare    holds a Doctor of Philosophy in Computer Engineering degree from Mumbai University, India in 2023. She is working as an assistant professor at K.J. Somaiya Institute of Technology, Mumbai, India. She has nearly 21 years of teaching experience. Her areas of expertise include data structures, algorithms, automata theory, soft computing, ML and DL, and bioinformatics. She has nearly 20 research publications and 2 collaborative research projects in the healthcare domain. She can be contacted at email: reena.l@somaiya.edu.







Dr. Jyoti Sunil More    pursued B.E. Computer Science Engg. (2003) from Shivaji University, M.Tech. in Computer Engineering (2006) from Dr. B.A.T.U., Lonere and Ph. D. in Computer Engg. From University of Mumbai in 2019. She has more than 20 years of Teaching experience and is currently affiliated to Fr. C. R. I. T, affiliated to Mumbai University. She has published several research papers in international conferences and journals. Her area of expertise are computer networks, databases, data mining, ML, data science, and AI. She can be contacted at email: jyoti.more@fcrit.ac.in.







Dr. Vaishali V. Sarbhukan (Bodade)    has completed her Ph.D. (IT) in 2019 from Mumbai University. She has completed her M.E. (Computer) and B.E. (CSE) in 2013 and 2002 from Mumbai University and Amaravati University respectively. Currently she is working as associate professor, Department of Information Technology, FCRIT, Vashi, Navi Mumbai. She has 19 years of teaching experience. She has published 26 papers in various scopus, SCI and international journal and 13 papers in national and international conferences. Her area of interest is security, ML, data science, and AI. She can be contacted at email: vaishali.bodade@fcrit.ac.in.







Mansing Rathod     Associate professor in the department of Information Technology, K.J. Somaiya Institute of Technology, Mumbai University and having 24 years' experience in teaching. Completed Ph.D. in image processing domain with research area i.e., Resolution Enhancement of Satellite image. He has authored over 34 peer reviewed conference and journal papers. He conducted several national level faculty development programs under AICTE. He can be contacted at email: rathodm@somaiya.edu.



Sarita Rathod     perusing a Doctor of Philosophy in Computer Engineering degree from Mumbai University, India. She is working as an assistant professor at K.J. Somaiya Institute of Technology, Mumbai, India. She has nearly 18 years of teaching experience. Her areas of expertise include data structures, ML and DL, and NLP. She has nearly 20 research publications. She can be contacted at email: sarita.r@somaiya.edu.



Dr. Sunita Patil     is a director at SVKM's NMIMS Shirpur, India. She is a member, Board of Studies in Computer Engineering, UoM. She received Ph.D. in Computer Engineering in the domain data mining, big data, and data science. She is having around 20 years of teaching and administrative experience. She has published her research work in various recognized National/International Journals and conferences. She has visited various international universities and organizations for attending conferences, knowledge sharing and exchange of information. Her passion is to bring in various outcome-based reforms in the field of academics contributing to the growth of society, nation and world at large. She can be contacted at email: spatil@somaiya.edu.