Explainable AI for Breast Cancer Risk Prediction: Evaluating the Accuarcy-Explainability Trade-off

Sona Barseghyan[1], Armine Babajanyan[1], Zhanna Balyan[1], Anna Asatryan[1], Sachin Kumar[1]
[1]Akian College of Science & Engineering, American University of Armenia, Yerevan, Armenia
sonabarseghyan04@gmail.com; armine.babajanyan.07@gmail.com; balyanzhanna04@gmail.com;
aasatryan04@gmail.com; s.kumar@aua.am

***Abstract*—The increasing application of artificial intelligence (AI) for impactful applications, such as healthcare, highlights the importance for models to not only make accurate predictions, but also to provide transparency in their decision making, such that humans can interpret their conclusions. This paper explores the integration of Explainable Artificial Intelligence (XAI) techniques in machine learning models for breast cancer detection using a real-world, large-scale dataset from Breast Cancer Surveillance Consortium (BCSC) with more than 1.5 68 million records. State-of-the-art classification models such as Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) were employed with special emphasis on decreasing the false negatives to reduce the likelihood of missed cancer. XAI approaches like SHAP, LIME, Layer-wise Relevance Propagation (LRP) were used to improve model interpretability. The rates of recall were high for the models, which is important in clinical screening, and this involved sacrificing precision, which is justified by the importance of early detection. Explainability analyses showed that biologically relevant risk factors that had been identified in clinical studies, including biopsy history, age, hormonal factors, and breast density, were the major contributors to model predictions. These findings highlight the potential of integrating XAI techniques into predictive healthcare systems, offering a path toward more transparent, reliable, and ethically sound AI-driven diagnostic support tools.**

***Index Terms*—Explainable AI, Breast Cancer Detection, SHAP, Machine Learning, Ensemble Learning, Model Interpretability.**

I. Introduction

Explainability is the ability of AI systems to provide a clear rationale for their actions and decisions, i.e., Explainable Artificial Intelligence can provide humans with explanations for its decisions or predictions. As illustrated in Figure 1, Explainable AI (XAI) overlaps with core AI domains like machine learning and deep learning. The growing trend of using AI in critical domains increases the need for transparency, trust, and ethical accountability [1]. Domains such as healthcare directly impacting people's lives rely heavily on explainability to minimize risks associated with hidden biases in prediction and classification models. For example, when diagnosing diseases with a high degree of accuracy, professionals need to be confident that the reasoning behind the result is sound, not just that the prediction is accurate, which is essential in maintaining trust, understanding, and transparency.

The transparency issues were raised in 2015, at a time when AI systems, particularly deep learning, were evolving rapidly [2]. The formal term XAI was prompted by a DARPA program launched in 2016 to support end users, such as intelligence analysts and autonomous system operators, who need to understand and trust AI decisions [3]. The DARPA program recognised the key challenge: the trade-off between Accuracy and Explainability, which is one of the important research concern. High-performance models such as deep neural networks were difficult to explain, while simpler, more transparent models often lacked accuracy (Figure 1).

The first applications of XAI focused on defense and national security, where human oversight of AI-generated decisions was critical [3]. As AI expanded into fields such as healthcare and finance, impacting diagnosis and fraud detection, the need for accountability grew. At the same time, concerns about the fairness and accountability of AI were growing globally. The introduction of GDPR in 2018 gave people in the EU the right to ask for explanations when decisions are made by automated systems, highlighting how important transparency had become [4]. This legal change combined with growing interest from researchers and policymakers, helped to bring XAI into the spotlight and encourage its wider adoption across different sectors.

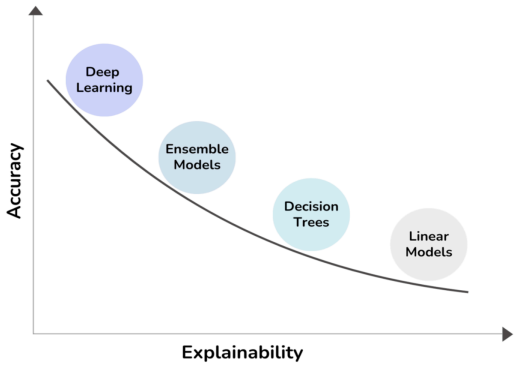

Fig. 1. Accuracy vs Explainability

II. Literature Review

Understanding how machines work has been central to designing them since their inception. While there has long been a focus on the design and construction of machines, there has also been an effort on the part of humans to gain full control and understanding of these creations. However, despite these efforts, gaining a comprehensive understanding

of how machines operate has proven more difficult than originally anticipated, especially as systems began to exhibit unexpected and unfamiliar behaviors that were not anticipated by researchers [6].

This challenge has persisted long before the formalization of explainability, as well as the development of artificial intelligence (AI), machine learning (ML), and deep learning (DL) as distinct fields of study [7]. As complex systems began to demonstrate behavior that was difficult to interpret, the necessity for transparency became more pressing [8]. In the late 1980s, the field of human-computer interaction began to address this issue by exploring ways to bridge the gap between machine operations and human understanding [9].

In their influential book Understanding Computers and Cognition, Winograd and Flores (1986) [10] argued that for a machine to be meaningful, it should align with the user's mental model of the world. While terms like "user-friendly" or "easy to learn" were commonly used, Winograd and Flores challenged this language, advocating for scientifically defined systems that humans could truly comprehend and rely on. This idea laid the foundation for later research on explainability.

Around the same time, rule-based expert systems emerged, using logic and clearly defined rules to represent knowledge. These systems improved transparency by simplifying the understanding of their processes [11]. However, as machine learning algorithms and neural networks evolved, these systems became increasingly accurate but also more complex, making them harder to understand. This complexity led to the "Black Box" problem, where machines provide accurate predictions but do not explain the reasoning behind them.

The first major breakthrough in explainability came with DARPA's Explainable Artificial Intelligence (XAI) program, launched in 2015. The initiative aimed to help users understand and manage intelligent systems more effectively. This marked the beginning of formal research into explainable AI, emphasizing not just model accuracy but also transparency. As a result, there was a growing demand for AI models to provide clear, understandable explanations for their decisions.

In the following years, researchers developed several tools and methods to make AI systems more transparent. One of the earliest and most prominent approaches was the Local Interpretable Model-Agnostic Explanation (LIME) by Ribeiro et al. [12]. This method allowed users to understand machine operations by providing explanations for individual predictions. Essentially, LIME created a simpler model around a specific prediction, helping users identify which features were most influential in the decision-making process.

Following LIME, Lundberg and Lee [13] introduced Shapley Additive Explanations (SHAP), based on Shapley values from coalitional game theory. SHAP assigns a fair score to each feature, assessing its importance for a given prediction. This method not only explains the decision-making process but also highlights the features that significantly impact the model's output.

While both LIME and SHAP were key developments in explainability research, they had a limitation: they provided local explanations but lacked a global understanding of the model's decision-making process. This gap in transparency made it difficult to fully trust AI systems. Christoph Molnar [14] explored this issue in his book Interpretable Machine Learning, distinguishing between local and global interpretability. He emphasized the importance of global transparency and further explored techniques like SHAP and LIME.

Additionally, the concept of global surrogate models emerged. These simpler models mimic the behavior of more complex models, helping humans understand AI systems better. Guidotti et al. [11] categorized surrogate models as part of the "Black Box Explanation Problem." Their work highlighted the value of replacing opaque models with globally transparent surrogate models, enabling users to understand the logic behind complex AI systems.

As research on explainability continued to grow, the need for organizing and categorizing XAI methods became clear. One of the first taxonomies was proposed by Adadi and Berrada [15], which classified XAI models based on complexity, scope, and model dependence. This framework laid the groundwork for future taxonomies, and many researchers, including Linardatos et al. [16] and Islam et al. [17], have refined these classifications over time. More recently, Martins et al. [18] developed a taxonomy based on stage, model dependence, and scope, emphasizing the trade-off between interpretability, human usability, and accuracy.

In conclusion, explainability research has significantly advanced, enhancing transparency, reliability, and trust in AI systems, and continues to evolve to ensure that end-users can interact with AI models effectively.

III. The Trade-off between Explainability and Accuracy

Explainable models—such as regression models and decision trees—typically exhibit lower predictive accuracy compared to black-box models like deep neural networks. However, their key advantage lies in interpretability, enabling stakeholders to understand and evaluate the rationale behind model decisions. This contrast presents a core challenge in AI model selection: the trade-off between explainability and accuracy [19].

This trade-off has become a focal point of debate within academic and professional communities. Researchers differ on whether to prioritize explainability or accuracy, often depending on the specific application and context of the AI system [20].

A. Viewpoints Advocating for Explainability

Many researchers advocate for explainability, asserting that interpretable models do not necessarily compromise performance. Candelon et al., for example, argue that models can be both accurate and transparent [21]. A study of 100 datasets found that in 70% of cases, explainable models achieved comparable predictive accuracy. These results emphasize the importance of considering the domain and goals of the system, as black-box models may introduce complexity without substantial performance gains [22].

Cynthia Rudin [23] takes a stronger stance, criticizing the use of black-box models followed by post hoc explanations. She calls for the development of inherently interpretable models—those that are transparent by design. Rudin argues that such ante-hoc models are not only applicable in practice but may also outperform black-box models. She further critiques post hoc methods for often yielding incomplete or misleading explanations that do not faithfully represent the model's logic, thereby reducing their trustworthiness and effectiveness.

B. Viewpoints Supporting a Balanced Perspective

Other scholars argue that both explainability and accuracy are essential and should be pursued simultaneously. Petkovic [24] challenges the notion that the two are mutually exclusive, emphasizing that accuracy alone is insufficient if users cannot understand or trust the model. Research shows that even accurate models may rely on irrelevant or flawed features, as revealed by XAI techniques. Petkovic advocates incorporating XAI throughout the development lifecycle to enhance transparency and user trust.

Crook et al. [25] propose a more nuanced view through their PET+ framework, which extends the traditional Performance-Explainability Trade-Off model by introducing a third factor: time, or development resources. They argue that the trade-off is context-dependent, shaped by resource constraints such as deadlines and computational costs. While explainable models can theoretically match black-box performance, achieving this often requires greater investment.

PET+ also stresses the importance of domain-specific considerations. In resource-constrained environments, simpler black-box models with basic explanations may be appropriate. Conversely, when time and resources are available, hybrid models that balance both aims should be favored. In high-stakes fields like healthcare, explainability may be prioritized over marginal performance gains, as errors can have serious, irreversible consequences [25].

IV. Related Work on Machine Learning Approaches to Breast Cancer Detection and the Role of Explainable AI

In this paper, a case study on breast cancer detection is conducted, exploring various machine learning methods that have been widely used for this purpose. This section focuses on the most commonly applied techniques, highlights existing research gaps, and discusses how integrating explainable artificial intelligence (XAI) can improve the interpretability and trustworthiness of predictive models.

One of the most prominent machine learning algorithms in breast cancer detection is Extreme Gradient Boosting (XGBoost). In a study by F. Silva-Aravena et al. from the Universidad Católica del Maule in Chile, XGBoost was used to train a model on a dataset of Indonesian women diagnosed or not diagnosed with breast cancer [5]. The algorithm combines multiple decision trees, where the final prediction is based on the aggregate output of all trees. It is designed to minimize overfitting, ensuring the model does not simply memorize training data but generalizes well to new inputs. In their case study, predictions over 50% were classified as positive, indicating the presence of breast cancer. XGBoost's high precision—the rate of correctly predicted positive cases—contributes to its widespread use. A similar approach was taken by Rahmanul Hoque et al., who noted XGBoost's resistance to feature multicollinearity and its ability to highlight key predictive features, achieving an accuracy of 94.74% [26].

Another powerful technique is the Support Vector Machine (SVM), which builds a decision boundary to separate different classes based on feature sets. In their research titled "Breast Cancer Diagnosis Using a Novel Parallel Support Vector Machine with Harris Hawks Optimization", Sultan Almotairi et al. proposed an enhanced model combining SVM with Harris Hawks Optimization. Trained on the Wisconsin Diagnostic Breast Cancer dataset, the model achieved 99.47% accuracy using an equilibration scaling technique [27]. The optimization process consists of exploration—searching for diverse hyperparameter combinations—and exploitation—refining the best-found configurations to maximize accuracy.

The third method is based on Artificial Neural Networks (ANNs)—computational models inspired by the structure and function of biological neurons. ANNs consist of layers of interconnected artificial neurons that process information through mathematical transformations. In Md Haris Uddin Sharif's study titled "Breast Cancer Detection Using Artificial Neural Networks", an ANN model was trained over 100 epochs, using ReLU activation in hidden layers and sigmoid activation in the output layer [28]. The model achieved a strong accuracy of 98.24% in predicting breast cancer cases.

Despite the strong results achieved by these methods, several important challenges persist. A major concern is the lack of transparency and interpretability, which limits clinical adoption. Models may arrive at correct predictions for the wrong reasons. For example, if the dataset shows that many women diagnosed with breast cancer are around age 50, a model might over-rely on the age feature, flagging any 50-year-old as high risk—even when other factors are more relevant. This underscores the need for XAI, which can explain how and why a model made a specific prediction.

To address these limitations, the following XAI techniques can be integrated:

1) SHAP (Shapley Additive Explanations): Assigns each feature a value representing its contribution to the model's prediction, helping radiologists understand the rationale behind a diagnosis.
2) LIME (Local Interpretable Model-Agnostic Explanations): Builds simpler, interpretable models that approximate the complex behavior of deep learning models, giving clinicians insight into the decision-making process.
3) Grad-CAM (Gradient-weighted Class Activation Mapping): Visually highlights regions of medical images that influenced the model's decision, allowing doctors to

assess whether the model focused on clinically relevant areas.

4) Grad-CAM (Gradient-weighted Class Activation Mapping): Visually highlights regions of medical images that influenced the model's decision, allowing doctors to assess whether the model focused on clinically relevant areas.

Incorporating these techniques can improve transparency, reliability, and ultimately the clinical usefulness of machine learning-based breast cancer detection models.

V. Methodology

In this section, for each of the discussed breast cancer detection models, a respective XAI method will be introduced to deal with transparency issues. Starting with the Extreme Gradient Boosting algorithm, the integration of SHAP (Shapley Additive Explanations) will be the best for achieving a deeper level of interpretability. Tree-based models such as XGBoost benefit from a model-specific implementation of SHAP (TreeSHAP), which enables exact and fast computation of Shapley values. Holzinger et al. in their work named "xxAI - Beyond Explainable AI" [29] , give the description of SHAP method as follows: the core idea is to explain the difference between the model's prediction for an individual case $f(x*)$ and a baseline value $e\phi$, typically the expected model output, as a sum of contributions from each feature:

$$f(x^*) = e\phi + \Sigma\phi(i) \tag{1}$$

where $\phi(i)$ is the Shapley value representing the contribution of feature i to the prediction. This makes it highly suitable for use in breast cancer detection models, where it is not only critical to predict accurately but also to explain why a given diagnosis has been made. In the case of Support Vector Machines (SVM), LIME (Local Interpretable Model-Agnostic Explanations) is adopted as the primary explainability technique. SVMs may sometimes transform input data into higher dimensions. However, this kind of transformations lack some logic and explainability for clinicians. So, LIME addresses this limitation by approximating the black-box model typically with a sparse linear regression—around a specific prediction. As described by Holzinger et al. [29], LIME addresses this interpretability gap by treating the SVM as a black-box function. It constructs a simple, interpretable surrogate model (typically a sparse linear model) that locally approximates the complex decision boundary of the SVM in the vicinity of the data point of interest. This is achieved by generating a neighborhood of samples around the original input and observing the SVM's predictions. These samples are then weighted by their proximity to the original point and used to fit the surrogate model. The resulting linear model provides insights into which features most influenced the prediction, offering clinicians a clear and intuitive understanding of the SVM's decision for a specific case. Last but not least, to interpret the predictions of Artificial Neural Networks (ANNs), the Layer-wise Relevance Propagation (LRP) method should be employed. As described by Holzinger et al. [29], this method redistributes the output prediction backward and assigns each neuron a relevance score. This way, experts can get some idea about the most important features contributing to prediction. LRP is not model-agnostic but specifically tailored for neural networks, enabling it to capture the structural intricacies of deep models. In the context of breast cancer detection, this enables clinicians to trace predictions back to specific biological or morphological markers in a patient's feature profile.

VI. Results and Discussion

A. Dataset Description and XAI Integration into the Implemented Algorithms

In the scope of this project the Extreme Gradient Boosting, Support Vector Machine and Artificial Neural Network algorithms were implemented with the integration of Explainable Artificial Intelligence techniques to introduce explainability to the models. Also the Dempster-Shafer algorithm was implemented which itself already contains explainability elements. The dataset for the analysis was sourced from the Breast Cancer Surveillance Consortium (BCSC), which is a collaborative network of six active breast imaging registries and two historic registries focused on research to assess and improve the delivery and quality of breast cancer screening and related outcomes in the United States [30]. Three related parts were retrieved and merged together to get the final dataset. The dataset includes risk factors related to breast cancer for individuals, like age, family history, hormone therapy, and other medical information linked to breast cancer risk. The resulting dataset has over 1,5 million of rows and 13 features. Some data preprocessing was done, including the elimination of observations with unknown values. Feature engineering was conducted to remove potential bias in the resulting predictions, such as performing one-hot encoding technique on the race column to reduce overfitting. As the target variable, the column containing breast cancer history information was taken, where it is indicated whether a certain person has had cancer history or not. It is worth mentioning, that while implementing the algorithms, a focus was put on optimizing the recall, which the proportion of the actual positive cases that the model correctly identifies. This is done to minimize the number of false negative cases, because predicting that an individual does not have cancer while the latter has it, can be vital, even fatal for that person.

B. XGBoost Algorithm Explained with SHAP

Starting from the Extreme Gradient Boosting(XGBoost) algorithm, it is worth mentioning that because of its nature of learning on its own previous mistakes, it performed quite well. The data was split into training and test sets, and the training of the model was performed. To get the most accurate and useful result, a wide range of hyperparameters were considered, and the combination of those hyperparameters that yielded the best result, was taken. After constructing a reasonable model, explainability was integrated to see whether the model actually works logically. The Shapley Additive Explanations(SHAP),

which provides consistent, mathematically grounded feature attributions, was integrated to the model. The method yields both local and global interpretations of the algorithm. Figure 3 shows the global explainability of the model, showing for the whole training set, which features contributed the most for the prediction.

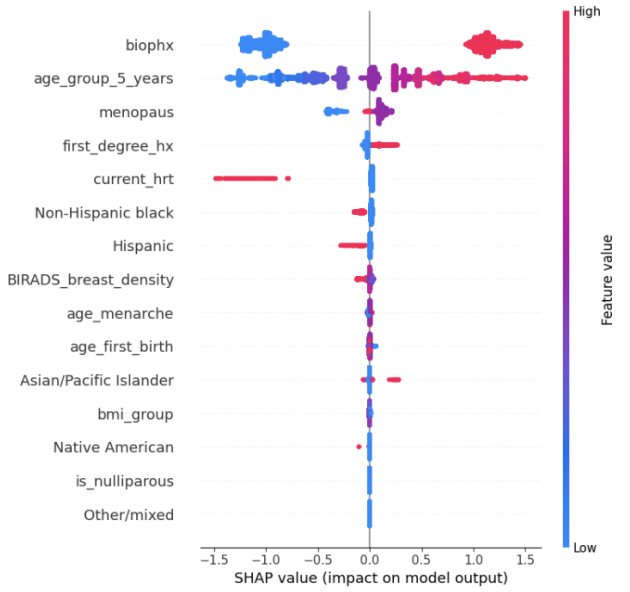

Fig. 2. Global Explainability with SHAP method

The SHAP method also gave some local explainability insights, showing for one random observation, which features contributed the most to the prediction. The result obtained in Figure 4 illustrates which features and with what exact values contribute the a certain observation being predicted as non-cancerous.

Fig. 3. Local Explainability with SHAP method

The integration of the explainable methods into XGBoost algorithm does not yield very promising results. The graphs conclude such features to be important such as race, particularly Non-Hispanic black feature, which is not very explainable and logical in this case. So, the XAI methods showed that even a reasonably accurate model has some logical drawbacks which can be disadvantageous in the clinical sphere.

C. SVM Explained with LIME

Following the experiments with XGBoost, the Support Vector Machine (SVM) algorithm was implemented with the primary objective of minimizing false negatives, which is critical in breast cancer detection. The model was trained using a radial basis function (RBF) kernel and class weight balancing to address class imbalance. Hyperparameters were optimized to improve recall, achieving a value of approximately 78%, while maintaining a strong ROC AUC score of 0.82. Although precision for the positive class remained low at 17%, this is a deliberate trade-off in screening applications, where identifying as many potential cancer cases as possible is prioritized over precision.

The Local Interpretable Model-Agnostic Explanations (LIME) approach was used to verify the credibility of the model predictions. LIME revealed the particular features that led to a classification on the individual level. As illustrated in Figure 4, the model process of decision closely follows the known clinical risk factors for breast cancer.

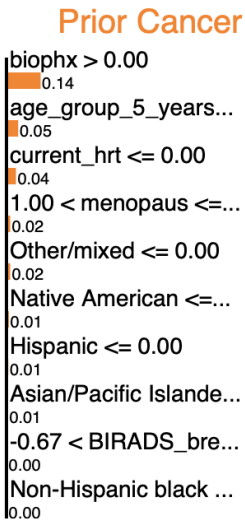

Fig. 4. LIME explanation for a sample prediction

Contributing factors were:

- **biophx (Biopsy History)**: History of prior biopsies dramatically raised the likelihood of cancer, as would be expected.
- **age_group_5_years**: Older age groups had higher cancer relative risk as described by epidemiological data.
- **menopausal status**: Postmenopausal status was also associated with high risk, supporting previous medical literature.
- **BIRADS Breast Density**: Lower BIRADS breast density was associated with lower cancer prediction, consistent with established diagnostic patterns.

Certain demographic variables were represented, i.e. ethnicity indicators; however their influence on the ultimate decision in this particular case remained negligible or neutral. Crucially, the model did not simply cherry-pick its decision based on soft demographic variables, either, but also relied heavily on clinically relevant factors such as biopsy history, age, hormonal factors, and breast density.

In addition, it is important to note that the model's performance metrics were consistently replicated on the test set, matching the validation results across all key indicators. This consistency demonstrates the model's strong generalization capability and suggests that neither the hyperparameter tuning

process nor the threshold adjustments introduced overfitting. The similar results across both datasets also confirm that the training, validation, and test partitions were representative of the overall data distribution, strengthening the reliability of the findings.

This is an interesting result, suggesting that, although there is a statistical trade-off for recall, the model's intrinsic logic for categorising high-risk cases is rooted in medically relevant reasoning. Such transparency supports its potential role as a preliminary screening tool, prompting further medical evaluation rather than serving as a definitive diagnostic system.

D. ANN Algorithm Explained with LRP

Next to bring the model of Artificial Neural Network into life, the dataset was first preprocessed by using imputation through the mean and handling missing values with that. After this step, the data was normalized by applying standard scaling technique. The data was then split into training and testing sets with an additional set further kept for validation using stratified split. Moreover, as the dataset was highly imbalanced, also the Synthetic Minority Oversampling technique (SMOTE) was applied to reach better generalization of the model on the testing and validation. The model was trained for over 30 epochs to reach a better accuracy. In the process, binary cross entropy was used as a loss function and Adam was used for optimization. After the model was trained, it was evaluated with the help of AUC score and confusion matrix values. As the model didn't reach a desired high accuracy at first, the classification threshold was fine-tuned to select one that allowed to maximize the recall while reaching acceptable precision.

As the aim of the study was to integrate Explainability into Machine Learning models to interpret the decisions they make and understand the reasoning that is hidden behind, the Layer-wise Relevance propagation(LRP) technique was used. The latter distributes a probability score across input values in different layers moving backwards to see which of those features influenced and contributed more to the model's decision. To make this true, the last activation layer of the Artificial Neural Network was removed and the LRP-ε rule was applied. The final output demonstrated which clinical features were the most efficient in their contribution to the model's decision.

The model placed significant weight on the patient's age group, a known risk factor for breast cancer. Additionally, the presence of a personal biopsy history (biophx) further elevated the risk prediction, aligning with clinical understanding that abnormal biopsy results are strong predictors of malignancy.

E. Algorithm Based on Demster-Shafer Theory

As a relatively new method that incorporates both good performance and explainability, we have implemented the model based on Dempster-Shafer Theory proposed by S. Peñafiel et al. in their paper "Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier" [31]. The theory requires the definition of Mass Assignment Functions (MAFs), which assign weights or probabilities to all subsets of potential outcomes based on the presence of specific facts within a decision-making context. In the context of breast cancer research, outcome can be represented as the following: If age interval is from 40-45, the model gives x weight to having cancer history, y weight to not having it and z weight to uncertainty.

Firstly, the model generates rules based on the features in the data and assigns random mass probabilities to each outcome: "History", "No History" and "Uncertainty". Afterwards, the model uses Gradient Descent algorithm to find the best combination of mass values. Each sample is classified according to the combination of the mass values of the rules that are applied for the sample. For example, in the context of breast cancer, an Asian woman aged 50, with an age of menarche of 14 and a age of menopause of 49 will be classified according to the combination of mass values provided by the rules that apply for the features that she possesses. If the mass of having cancer history is higher, she will be classified as a patient with a history of cancer and vice versa.

The algorithm not only effectively classifies the samples but is also very explainable. At the end of the training the resulting rules are printed, and the weights of each rule's outcome can be seen and interpreted. The results of training on breast cancer data can be seen in TABLE I. As the number of rules for the data was very large, the table shows only the top 10 rules that have the highest mass for having a cancer history.

TABLE I
RULES AND MASS WEIGHTS GENERATED BY THE DEMPSTER-SHAFER CLASSIFIER

Rule	No History	History	Uncertainty
If age_group_5_years == 13	0.037	0.707	0.256
If biophx == 1	0.125	0.527	0.348
If current_hrt == 0	0.189	0.521	0.291
If age_group_5_years == 12	0.152	0.516	0.332
If age_group_5_years == 10	0.132	0.388	0.480
If age_group_5_years == 11	0.232	0.359	0.409
If menopaus == 2	0.255	0.350	0.395
If Non-Hispanic black == 0.0	0.228	0.319	0.453
If Other/mixed == 0.0	0.296	0.319	0.385
If Native American == 0.0	0.324	0.310	0.366

According to the table, high age groups show a high mass in the likelihood of cancer, which is consistent with general knowledge about cancer. Also, the fact that having had a biopsy also increases the likelihood of having cancer speaks about the model catching on a relevant feature which aligns with common sense. The values of having a history of breast cancer in the table might seem small, but considering the fact that the majority of women who have the disease will have a combination of factors that possess a high mass of history, the overall chance of having the disease will become high, leading to the model correctly classifying it.

Coming to the performance metrics, the model has shown the following results from TABLE II. While the performance is not perfect, it is still quite good for the given dataset in comparison with other models.

TABLE II
PERFORMANCE OF THE DEMPSTER-SHAFER CLASSIFIER

Metric	Value
Accuracy	0.7571
ROC AUC	0.8304
F1 Score	0.7457
Precision	0.7794
Recall	0.7148

Overall, we can conclude that the model is quite good combination of accuracy and explainability, which means that it is possible to get the maximum from the explainability-accuracy trade-off.

VII. CONCLUSION

In this work, we investigated the incorporation of Explainable Artificial Intelligence (XAI) methods in machine learning models for breast cancer detection, with the purpose of trading off predictive ability and transparency of models. In addition to standard model performance metrics, use of XAI techniques, including SHAP, LIME, and Layer-wise Relevance Propagation (LRP), is also insightful for understanding rationales behind model predictions and identifying key risk factors affecting the outcome. Not only did this emphasize the issue of accuracy against interpretability, but it also underscored the need to maintain transparency in AI systems used in healthcare, where the effects of black-box decisions can have profound impacts.

We considered a variety of models such as XGBoost, Support Vector Machines (SVM), Artificial Neural Networks(ANN), and the Dempster-Shafer framework. XGBoost achieved the best trade-off between accuracy and ROC AUC and SVM achieved a good trade-off with highly interpretable local explanation with LIME. The neural network performed worse than XGBoost but was able to leverage LRP explanations to identify features of importance. The Dempster-Shafer method resulted in a system that provided balance and completeness and was indicated as worthy of further investigation in this field.

Nevertheless, some limitations were observed in the present study. The size of the sample was restricted to 50000 rows for computational reasons that may limit the accuracy of the model as discussed above. In addition, the study largely utilized available clinical characteristics and did not consider more elaborate biological and genetic information. Future studies could improve predictors and be more clinically translation by addressing these limitations, using larger datasets and incorporating more advanced bioinformatics functionalities.

In the end, this work highlights the need for AI systems to be not only accurate but also transparent and ethically grounded. This work is a step towards building a foundation for responsible AI adoption in healthcare, in which interpretable and trustworthy models facilitate clinicians' decisions by adhering to fairness and accountability.

VIII. INTERACTIVE EXPLANATIONS FOR FURTHER POSSIBILITIES IN XAI

Most widely used XAI systems today are static, providing fixed, auto-generated explanations without user adaptability [32]. These one-time outputs often include feature importance scores (e.g., SHAP values) or visualizations such as decision trees and feature attribution heatmaps. While static explanations offer valuable insights, they lack flexibility and personalization, making them less suitable for broader or more user-specific applications [32].

This limitation has led to growing interest in interactive explanations. Unlike static systems, interactive XAI allows users to engage with and refine explanations based on their feedback. These systems adapt to individual needs, offering more tailored and comprehensible insights. They are particularly well-suited for contexts such as AI-assisted research and decision-making.

Mindlin et al. compared static and dialogic explanation systems to assess their impact on user understanding [32]. Their study found that dialogue-based approaches led to significantly higher comprehension. Participants who better understood the model tended to ask more specific, feature-related questions and engaged more deeply with the system, highlighting the role of active exploration in enhancing interpretability.

Despite this promise, the field of interactive XAI remains underexplored. Bertrand et al. emphasize the lack of structure and taxonomy in current research. Their study organizes 48 papers into three categories: Selective, Mutable, and Dialogic systems [33]. Selective systems enable users to focus on particular components of a model or explanation. Mutable systems allow manipulation of inputs or parameters to observe changes in outputs. Dialogic systems facilitate back-and-forth interactions between user and system. Notably, most research has focused on mutable systems, while dialogic approaches have received comparatively little attention—leaving substantial room for further investigation.

Bertrand et al. also found that interactive features increase the perceived usefulness of explanations and enhance collaborative performance between AI and humans [33]. They also observed a rise in the amount of time users spend interacting with XAI systems, suggesting higher engagement levels.

To address the existing research gap, future work should focus on systematically evaluating interactive XAI systems against static counterparts. Comparative studies are needed to assess not only differences in user comprehension and engagement but also their practical impact on decision-making and system effectiveness.

REFERENCES

[1] EDPS (European Data Protection Supervisor), 2023, $https://www.edps.europa.eu/data-protection/our-work/publications/annual-activity-reports/2024-04-19-annual-activity-report-2023_enaccessedon11Apr2025$.

[2] Dwivedi, Y. K., Sharma, A., Rana, N. P., Giannakis, M., Goel, P., & Dutot, V. (2023). Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions. Technological Forecasting and Social Change, 192, 122579.

[3] D. Gunning, E. S. Vorm, J. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," *Applied AI Letters*, vol. 2, 2021. [Online]. Available: https://www.researchgate.net/publication/356781652_DARPA_'s_explainable_AI_XAI_program_A_retrospective

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, & D. Pedreschi, "A survey of methods for explaining black box models". ACM Computing Surveys, 2018, 51(5), 1–42. https://doi.org/10.1145/3236009

[5] F. Silva-Aravena, H. Núñez Delafuente, J. H. Gutiérrez-Bahamondes, and J. Morales, "A Hybrid Algorithm of ML and XAI to Prevent Breast Cancer: A Strategy to Support Decision Making," *Cancers*, vol. 15, no. 9, p. 2443, 2023. [Online]. Available: https://doi.org/10.3390/cancers15092443

[6] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos et al. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 56, 3005–3054 (2023). https://doi.org/10.1007/s10462-022-10246-w

[7] S. Razavi. "Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling." Environmental Modelling and Software 144 (2021): 105159.

[8] N. Balasubramaniam et al. "Transparency and explainability of AI systems: From ethical guidelines to requirements." Information and Software Technology 159 (2023): 107197.

[9] J. H. Gerlach, and F.-Y. Kuo, Understanding Human-Computer Interaction for Information Systems Design. MIS Quarterly, 1991, 15(4), 527–549. https://doi.org/10.2307/249456

[10] T. Winograd and F. Flores, Understanding Computers and Cognition, Norwood, N.J.: Ablex Publishing Corporation, 1986, 207 pp.

[11] 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 (CSUR), vol. 51, no. 5, pp. 1–42, Aug. 2018.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[14] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[15] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, 2021. [Online]. Available: https://doi.org/10.3390/e23010018

[17] M. U. Islam, M. M. Mottalib, M. Hassan, Z. Alam, S. M. Zobaed, and M. F. Rabby, "The Past, Present, and Prospective Future of XAI: A Comprehensive Review," in *Explainable AI: Foundations, Developments, Prospects and Challenges*, 2022.

[18] T. Martins, A. M. de Almeida, E. Cardoso, and L. Nunes, "Explainable Artificial Intelligence (XAI): A Systematic Literature Review on Taxonomies and Applications in Finance," *IEEE Access*, vol. 12, pp. 3347028, Jan. 2024.

[19] V. Hassija, V. Chamola, and A. Mahapatra. "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn Comput 16, 45–74 (2024).

[20] V van der, N. Sabine, et al. "Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries." Journal of the American Medical Informatics Association 28.10 (2021): 2128-2138.

[21] F. Candelon, T. Evgeniou, and D. Martens, "AI can be both accurate and transparent," *Harvard Business Review*, May 12, 2023. [Online]. Available: https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent

[22] A. Budhkar, S. Qianqian, S. Jing, and Z. Xuhong. "Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics." Computational and Structural Biotechnology Journal (2025).

[23] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.

[24] D. Petkovic, "It is not 'accuracy vs. explainability' – we need both for trustworthy AI systems," *arXiv preprint arXiv:2212.11136*, Dec. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2212.11136

[25] B. Crook, M. Schluter, and T. Speith, "Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)," *arXiv preprint arXiv:2307.14239*, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.14239

[26] R. Hoque, S. Das, M. Hoque, and E. Haque, "Breast Cancer Classification using XGBoost," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1985–1994, 2024. [Online]. Available: https://doi.org/10.30574/wjarr.2024.21.2.0625

[27] S. Almotairi, E. Badr, M. Abdul Salam, and H. Ahmed, "Breast Cancer Diagnosis Using a Novel Parallel Support Vector Machine with Harris Hawks Optimization," *Mathematics*, vol. 11, no. 14, p. 3251, 2023. [Online]. Available: https://doi.org/10.3390/math11143251

[28] M. H. U. Sharif, "Breast Cancer Detection using Artificial Neural Networks," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 9, no. 10, Oct. 2021. [Online]. Available: https://www.researchgate.net/publication/355982962_Breast_Cancer_Detection_using_Artificial_Neural_Networks

[29] A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., "xxAI - Beyond Explainable AI: International Workshop Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers." 2020.

[30] BCSC Research, "Risk Factor Dataset," *BCSC Research Datasets*, [Online]. Available: https://www.bcsc-research.org/index.php/datasets/rf/risk-factor-dataset-download?form_success=1&entry=67dc5da796402615920092#form1575

[31] S. Peñafiel, N. Baloian, H. Sanson, and J. A. Pino, "Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier," *Expert Systems with Applications*, vol. 148, p. 113262, 2020.

[32] D. Mindlin, A. S. Robrecht, M. Morasch, and P. Cimiano, "Measuring user understanding in dialogue-based XAI systems," *arXiv preprint arXiv:2408.06960v1*, Aug. 2024. [Online]. Available: https://arxiv.org/html/2408.06960v1

[33] A. Bertrand, T. Viard, R. Belloum, J. R. Eagan, and W. Maxwell, "On selective, mutable and dialogic XAI: a review of what users say about different types of interactive explanations," *HAL preprint hal-04115961v1*, Jun. 2023. [Online]. Available: https://hal-lara.archives-ouvertes.fr/DIVA/hal-04115961v1