A Survey Exploring the Application of SHAP in Law, Cybersecurity, Healthcare, and Finance

Mackenzie Chase
Department of Computer Science
University of New Brunswick
Fredericton, Canada
wlm.chase@unb.ca

Arthkumar Rakeshkumar Patel
Department of Computer Science
University of New Brunswick
Fredericton, Canada
arth.patel@unb.ca

Tianna-Lee Salmon
Department of Computer Science
University of New Brunswick
Fredericton, Canada
tia.salmon@unb.ca

Luis Eduardo Lazo Vera
Department of Computer Science
University of New Brunswick
Fredericton, Canada
lazo.vluis@unb.ca

***Abstract*—Artificial intelligence is advancing rapidly, with applications in sectors such as healthcare, finance, law and cybersecurity. However, many artificial intelligence models are considered a "black-box" due to their lack of transparency. This means that it is difficult for users to understand the factors that influence a model's decision-making process. This lack of clarity is especially concerning in high-stakes situations that involve life-threatening decisions, privacy or regulatory compliance. In these areas, transparency is not just desirable; it is essential. As artificial intelligence continues to be employed in various industries, ensuring clarity and openness is key to maintaining trust as well as adhering to ethical and regulatory standards. This survey examines the use of the SHAP explainable AI technique across various sectors.**

***Index Terms*—Explainable AI, XAI, SHAP, Law, Cybersecurity, Healthcare, Finance**

I. Introduction

How the human brain works and makes decisions based on provided information is an open question in the field of Neuroscience. The human brain contains approximately 10^{11} neurons connected in a holistic manner, defining a complex biological system [1]. Artificial Intelligence (AI) is the science of engineering intelligent machines that can simulate human cognitive processes [2]. Certain AI models are naturally *interpretable*. For example, understanding the inner workings of linear regression models and decision trees is relatively straightforward because their predictions are based on clear, explainable rules and relationships between variables. Interpretable models allow users to trace the logic behind their predictions, making it easy to understand how these decisions are made. On the other hand, "black-box" models, such as deep learning models, are more complex. Neural networks, for example, which are used for tasks such as image recognition, involve multiple layers of calculations and transformations. While it is possible to assess whether a prediction is accurate, understanding how the model arrived at that prediction is much more challenging. The underlying processes are so intricate that it becomes difficult to explain the reasoning behind each decision, making optimization and trust-building more difficult.

Explainable Artificial Intelligence (XAI) aims to improve the transparency of AI models. Understanding how these models make predictions and identifying which features contribute to their decisions is essential for optimizing performance and improving explainability [3]. The primary goal of XAI is to facilitate human interaction with AI systems, which is especially critical in fields such as healthcare [4] and law [5], where ethical considerations, privacy, fairness and safety play a crucial role in the responsible use of AI [6]. Shapley Additive Explanations (SHAP) is a widely used XAI technique, known for its ability to explain feature importance. SHAP provides both local explanations (detailing how each feature influences individual predictions) and global explanations (showing which features are most important across the entire model) [7].

Previous surveys have examined XAI from various angles. Several studies offer their opinions on XAI or propose solutions for improvement, providing either a general or domain-specific view on XAI. Bibal et al. [8] discuss the varying explainability requirements for legal AI models, emphasizing how these requirements depend on the legal context, ranging from the need for provide basic feature information to comprehensive model explanations. Benedetto et al. [9] explore how court judgment prediction can be improved using XAI and propose an explainable Natural Language Processing (NLP) technique (Legal NER) to achieve this goal. State et al. [10] examine the importance of model explanations in the context of the GDPR, highlighting challenges with current XAI techniques in the legal domain. The authors note that the adequacy of existing explanations depends on the specific legal problem at hand, with some explanations failing to meet the required legal standards for transparency. Brożek et al. [11] challenge

the conventional view of the black-box problem, arguing that explainability in the legal context should be evaluated based on psychological criteria, such as understanding and trust, rather than solely on technical explanations. The authors emphasize the need for sector-specific approaches to explainability, as each sector may have unique needs, a point similarly made by Bibal et al. [8].

In addition to opinion surveys, various studies propose sector-specific taxonomies of XAI. Zhang et al. [12] explore XAI techniques in the cybersecurity domain, compiling a wide range of methods, from common techniques like SHAP and LIME to less common ones such as gradient-based and visualization tools. The authors propose a taxonomy that includes the XAI method, cyberattack type, AI model, XAI type and scope. Černevičienė and Kabašinskas [13] examine XAI applications in the financial sector and suggest a taxonomy that includes the XAI method, AI model, application, XAI type and scope. Gupta and Seeja [14] focus on XAI in healthcare, proposing a taxonomy that includes the XAI method, XAI type and scope. Zhang et al. [15] discuss XAI in auditing and present a general taxonomy based on research in the auditing field. This taxonomy includes the XAI method, description, function, scope, advantages and disadvantages. Kute et al. [16] explore the interpretability of AI methods used for detecting money laundering, proposing a taxonomy that includes the AI methods used, whether the model is interpretable and the data type. Hacker et al. [17] discuss XAI in contract and tort law and propose a taxonomy explaining how different models are suitable for various types of explainability methods, such as transparency and post-hoc interpretability, with a focus on simulatability, decomposability and local explanations for legal decision-making processes. Górski et al. [18] analyze XAI in the tax domain, focusing on the comprehensibility and assessability of XAI methods. Richmond et al. [19] explore XAI in law, proposing a taxonomy that includes the XAI method, purpose and data type.

Many of the existing XAI taxonomies have overlapping features, leading to studies focused on merging these features into a more comprehensive taxonomy. Speith [20] highlighted that the range of explainability methods is too vast to be effectively captured in a single, pragmatically useful taxonomy. In other words, it is impractical to include all potential dimensions within one taxonomy, as it would be overwhelming to navigate. As a result, Speith focuses on evaluating major XAI taxonomies and proposes a final version that includes commonly used dimensions such as stage, scope, and output format. However, even this proposed taxonomy remains too complex to efficiently compare explainability methods. In addition, the final taxonomy does not address the contextual importance of an explainability method. Factors such as the AI model used, the input data type, the use case and the sector all play crucial roles in determining which explainability method to apply.

While existing XAI taxonomies are useful for exploring various techniques, their broad scope often makes them difficult to navigate and apply effectively. Some taxonomies, such as those by Zhang et al. [12], Černevičienė and Kabašinskas [13], and Kute et al. [16], consider the AI model and use case in their classification, but many others do not, leading to questions about the context in which each XAI technique is applied. This context is crucial when deciding on the appropriate XAI technique for a given scenario.

Furthermore, while there have been efforts to combine existing taxonomies into a single, comprehensive framework [20], there is still a significant gap in developing a unified taxonomy that spans multiple sectors. This taxonomy would not only facilitate comparisons of XAI methods across sectors, but also provide a clearer understanding of how XAI techniques apply to specific use cases. This survey proposes a more comprehensive taxonomy, focused on SHAP in healthcare, law, finance and cybersecurity. This taxonomy also includes AI models and use cases within their respective sectors. This approach contrasts with previous studies which either isolated the XAI techniques from their sectors or overlooked the importance of the deeper context within which the techniques are applied. As emphasized by Bibal et al. [8] and Zhang et al. [12], interdisciplinary exploration of XAI is crucial, especially since each sector faces unique challenges that require tailored approaches to explainability.

This paper is structured as follows: we begin with background information on popular XAI techniques, followed by a review and comparison of related works in the field. Next, we present our methodology and the taxonomy of SHAP across various sectors, concluding with a discussion of our findings.

II. Background

A. GradCAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a visual method that allows us to reveal the underlying features that contributed to making predictions in a Convolutional Neural Network (CNN) for image recognition [21]. The outcome from this method will be a heat map that overlaps with the input image and highlight the areas in the image that have been considered in the model to make that prediction. The aim of this approach is that it works in last layer of the CNN, and the model does not need to be retrain or performing any changes in the architecture [21]. The application of the CNN in areas of image classification and object detection, are widely discuses in the works of [22] and [23], those model use CNN to perform the classification. However, the model can not explain how those classifications are performed nor which areas in the image have been considered in the classification. Grad-CAM can highlight the areas of the image that is important in the classification. This is very important because, it can help in order to optimize the model, we can focuses in the areas that have significant relevance to perform better predictions and eliminate unnecessary areas from the data set without needing to retrain the model [21].

B. LIME

Riberio et al. [24] first introduced Local Interpretable Model-Agnostic Explanations (LIME) in 2016. LIME can

work on any classifier (model-agnostic). It operates locally in a model as opposed to globally, assessing individual predictions by making local linear approximations. They conducted experiments with image data, tabular data, and text-based data. In one particular experiment, they assessed if pieces of text had elements of Christianity or Atheism. The model correctly predicted excerpts provided along with explanations based on keywords into the appropriate class. However, some of the explanations had no connection to either class such as the word "Posting". They acknowledge that end users may not be able to look at each individual explanation for each individual prediction. This can be cumbersome in large models, which is why they further propose SP-LIME, or Submodular Picks – LIME, which enables a selection of explanations throughout a model to give a better global understanding of the model.

C. SHAP

SHAP represents a unified framework for providing interpretations to model predictions [7]. It assigns an important value to each feature, based on which it can identify the contribution of each feature in the response. It combines the concept of game theory (Shapley values) with local explanation methods to provide grounded feature attribution that satisfies three desirable properties: local accuracy, missingness, and consistency. SHAP combines six different methods for explaining model predictions into one framework, making it both computationally faster and more aligned with how humans understand model outputs [7]. Mosca et al. [25] provide a thorough review of SHAP based methods for interpreting Natural Language processing models, by examining 41 approaches they identified several key research directions, including adaptations for different input types and improvements in computational efficiency. Their research assesses these algorithms based on their implementation feasibility and applicability for text data, making specific suggestions for various NLP applications. For example, they propose HEDGE for feature attribution in large texts and Neuron Shapley for detecting biased neurons. The authors also emphasize significant shortcomings in current SHAP-based techniques, specifically their limited application to sequence-to-sequence tasks and computational issues with huge inputs, outlining critical paths for future research in NLP model interpretability.

D. TCAV

Testing with Concept Activation Vectors (TCAV), first introduced in [26], explore the possibilities in using CAVs as a method to quantify the importance of features in image classification and their possible use in medical applications. Their focus with TCAVs was to achieve accessibility, customization, plug-in readiness, and global quantification. TCAVs use concept data, such as patterns (stripes) as well as random data. The concept data is trained on the model first, recording the activations of each layer, followed by the random data. These activations in the layers are gathered to then create a directional derivative which are generally normal to the random data to determine which areas in the model is sensitive to the given concept. The authors acknowledge that the directional derivative or vectors can become less orthogonal when differentiating between similar concepts (black hair, brown hair). To ensure methodological rigor, 500 training iterations at minimum are advised for consistency verification. The framework includes comparative analysis with alternative models, with particular emphasis on saliency - the identification of the most important or prominent features. Finally, the researchers conducted an experiment for predicting diabetic retinopathy with strong results, however often over estimated the severity of diagnostics. With some human intervention, TVACs can accelerate workflows especially in the medical sector.

III. Related Works

Oliveira et al. [27] developed four models to predict inmate behavior and applied SHAP to understand which features were most relevant in predicting misconduct during detention. The models were built using the Survey of Inmates in State and Federal Correctional Facilities (SISFCF) dataset which includes inmates' personal history, such as their number of offenses, sentences, and drug history. The four models used in this study included Decision Trees, Neural Networks, Support Vector Machine (SVM) and Random Forest. Decision Trees are intrinsically interpretable (ad-hoc), meaning that SHAP only needed to be applied to the three remaining models. The study found that Random Forest performed the best, achieving an F1-score of 69% for the "Yes" class and 71% for the "No" class. The results of applying SHAP to the Random Forest model were further analyzed. The study revealed that the feature "Intentionally Injured" had the highest impact on predicting inmate misconduct. This feature referred to inmates who had been intentionally injured since being admitted to prison. A major limitation noted was SHAP's assumption of feature independence, and the study suggested exploring the Causal Shapley Values approach to overcome this limitation.

Zhang et al. [14] compared SHAP and LIME in auditing models, emphasizing the importance of adhering to auditing evidence standards, which makes XAI especially crucial in this domain. They developed an XGBoost model to predict the risk of material restatement which refers to the revision of previously issued financial statements due to errors or misstatements. Both SHAP and LIME were applied to the model to identify the features most likely to contribute to material misstatements. The study found that both SHAP and LIME performed similarly, although LIME struggled to identify important features when raw financial statement variables were used.

Černevičienė and Kabašinskas [27] conducted a systematic literature review (SLR) to explore XAI techniques in finance. They developed a taxonomy that categorized XAI techniques based on their type (local or global) and whether they were post-hoc or ante-hoc.

Zhang et al. [15] explored XAI techniques used in cybersecurity, identifying whether each technique was local, global, model-specific, model-agnostic, post-hoc, or ante-hoc, and whether the output was text-based, visual, argument-based, or

model-based. The taxonomy was organized based on different types of cyber-attack such as malware and botnets. A wide range of AI techniques were studied, including common ones like SHAP, Grad-CAM, and LIME, as well as less common techniques such as gradient, heatmaps and Fraud Memory. Although the taxonomy was informative, the wide range of XAI techniques made it difficult to determine which XAI method was most suitable for each use case.

Górski et al. [28] use Grad-CAM to examine the text-processing of legal documents. The study focuses on how the choice of embeddings affects downstream processing by implementing a pipeline containing an embedder, a CNN classification model and a metric-based evaluator. The CNN was trained using a PTSD dataset which was used to classify the rhetorical roles of a sentence as well as a Statutory Interpretation - Identifying Particular (SIIP) dataset which was used to classify sentences into four categories based on their usefulness to a legal provision. The PTSD dataset was taken from the U.S. Board of Veterans' Appeals (BVA) and contains adjudicated disability claims by veterans for service-related PTSD. The SIIP dataset is taken from the United States Code (U.S.C.) Title 5, Section 552a(a)(4) (5 U.S.C. § 552a(a)(4)) which includes text from court decisions related to the Privacy Act of 1974. Each sentence in the dataset is classified based on its usefulness in interpreting 5 U.S.C. § 552a(a)(4) under four categories: High Value, Certain Value, Potential Value and No Value. In order to compare Grad-CAMs output for two models, Gorski et al. propose two metrics. The first metric is "Fraction of elements above relative threshold, t" (F(t)) which measures how the CNN spreads its attention over the words in the input. It is defined as a number of elements in a vector that are larger than the relative threshold multiplied by the maximum vector value divided by the length of this vector. The second metric proposed is "Intersection over union with relative thresholds t1 and t2" (I (v1, v2,t1,t2)) which compares predictions of two given models on the same input sentence. It takes as arguments two Grad-CAM heatmaps (v1 and v2), binarizes them using relative thresholds (t1 and t2) and finally calculates standard intersection over union. It quantifies the relative overlap of words considered important for the prediction by each of two models. The architecture used in the study consists of four modules: the preprocessing module, the embedding module (based on BERT, word2vec and Law2vec), the classification module (the CNN) and the visualization module. Using Grad-CAM, a class activation map was generated showing how strongly each token influenced the CNN's prediction.

Kute et al. [16] develop a CNN model to predict suspicious transactions and employ SHAP to explain and validate the model's behaviour. The authors explain the need for XAI in this area due to the high FP rate of 98% in relation to transactions being flagged as suspicious. By employing XAI techniques, there is more transparency into why the model is falsely flagging a certain transaction as suspicious. With this insight, the model can be adapted to reduce the false positive rate. Another major issue of employing non-interpretable AI models is the lack of transparency as it relates to data privacy and ethical usage of the data. This is especially detrimental as regulations such as GDPR and CCPR come into effect. To address this, Kute et al. implement a Conv1D CNN, along with three other models for comparison (RF, XGBoost and SVM). However, SHAP was only applied to the best-performing model which turned out to be the CNN. All models were trained on synthetic data which was outlined as a limitation in the study. The synthetic dataset contained customers, accounts, legitimate transactions, and suspicious transactions. However, to increase realism, demographic information from the Australian Bureau of Statistics was used to generate customer profiles. For brevity, only the CNN model will be explained in detail along with the results of SHAP being applied to this model. The CNN model consists of five layers: Conv1D, batch normalization, flattening, dropout, and a dense layer. The Conv1D layer was chosen to work with the tabular dataset, convolving the layer input over a single dimension. The batch normalization layer normalizes the output of each layer to maintain a mean close to 0 and standard deviation close to 1. The dropout layer randomly sets a fraction of input units to zero during training, which helps prevent overfitting and improves model generalization. It is only active during training. The flatten layer converts multidimensional input into a one-dimensional array, preparing the data for the next fully connected layer. The dense layer connects every neuron from the previous layer and performs a weighted sum followed by an activation function, typically ReLU for hidden layers and sigmoid for output layers in classification tasks. After hyperparameter tuning, the optimal results were shown to be with 100 epochs and a batch size of 32. SHAP interpretation can be both local and global. However, the study focused on the local interpretations, focusing on the feature importance. Analyzing the SHAP force plot for the CNN prediction of a suspicious transaction, the key features were found to be credit, transaction amount, transaction description, KYC state, and transaction currency. In the SHAP force plot for the CNN prediction of a legitimate transaction, the key features were found to be transaction description, transaction location code, transaction location type, transaction currency, and transaction subtype. Based on analyzing the details of the record associated with both these predictions, the authors agreed with the analysis.

Luo et al. [29] use both SHAP and LIME to provide local explanations in Legal Data. The goal is to examine the correctness of these explanations using faithfulness and plausibility as metrics of evaluation. Three datasets were used. The first dataset is related to personal injury negotiation (PIN), the second to trademark confusion (TC) and the third to worker classification (WC). The model used was an XGBoost gradient-boosted decision tree trained with binary cross-entropy loss. The model was trained on each dataset and the results of SHAP and LIME was compared. SHAP was shown to perform best on the WC dataset while LIME performed best on the PIN dataset. In evaluating the performance of the XAI methods, both were shown to have inconsistent faithfulness and SHAP was shown to have better plausibility.

Saleh et al. [30] examine the importance of explainability in machine learning models through a case study for the biomedical domain, focusing on the strengths and weaknesses of SHAP and LIME techniques. Based on their findings, these strategies' effectiveness is significantly influenced by feature collinearity in the dataset and the choice of ML model. In their examination, SHAP demonstrated significant computational cost, especially when working with datasets that have a lot of features.

The primary conclusion derived in [31] is the imperative need for XAI in healthcare because of the high-stakes nature of this sector and the need for trust, transparency, and accountability in AI-driven decision-making. The authors conducted a systematic literature review of XAI focused on their applications in healthcare. They explored various ML and DL models and categorized XAI methodologies into feature-oriented, global, concept models, surrogate models, local pixel-based, and human-centric approaches. These XAI methods, including SHAP, LIME, CAM, Grad-CAM, and LRP, are applied for disease diagnosis, medical image interpretation, and understanding model reasoning. They highlight the lack of standardized evaluation for explanations, computational costs, the trade-off between performance and transparency, and the need for user-centric and contextually relevant explanations.

[32] provide an overview of XAI methods, categorizing them as perceptive and via mathematical structures, and how they are used within the health field. Their work involved a review of a wide set of research papers, and analyzed their interpretability. The authors find that even as there is increased research interest in XAI based on the need for transparency and accountability, particularly across fields such as medicine, there remains a pending challenge with regards to solving interpretability issues with black-box systems. Most studies have proceeded under the assumption of interpretability without sufficient human validation, whereas uniform adoption of evaluation criteria continues to be absent across research. They reiterate the potential XAI has in healthcare and the need to create trust in such a critical infrastructure. The lack of standard evaluation of XAI methods in clinical use, and the potential for generated explanations to be unreliable, manipulable, or based on noisy or incomplete data leaves room for improvement for these models.

[33] reviewed XAI techniques that are being applied in medical imaging, namely X-Rays, ultrasounds, and CT scans, and electronic health records (EHR). They proposed to categorize XAI techniques along the dimensions of interpretation type, model specificity, explanation scope, and explanation form. The paper examined Grad-CAM, which is effective in locating important regions in saliency maps but has limitations in model accuracy validation. LIME, as a model-agnostic tool for local visual explanations, suffers from poor repeatability. SHAP performs well for feature analysis, however time-consuming. TCAVs are capable of providing high-level concept-based explanations, yet their use in medical imaging must be researched further for more robust results.

The authors of [34] centers on reviewing LIME and SHAP. The study explores their practical implementation through case studies. The first case, concerned heart stroke risk forecasting, utilizes LIME to provide local interpretability for classification models. The methodology involved data partitioning, preprocessing, feature selection via Single Spectrum Analysis, and subsequent LIME analysis to understand the influential features in individual predictions. The key finding is LIME performed more accurately with gradient descent over the other approaches. They further highlight SHAP performs best in terms of interpretability with RF, Ada boost and Gradient boost. The second case study did not include investigations in specific XAI techniques, however explored DNN for a computer-aided diagnosis. The final case study explores XAI for Earlier Warning Score (EWS) for predicting critical illnesses. This approach does not utilize specific XAI techniques, however, highlights strong accuracy of XAI-driven model providing physicians justifications in predictions from identifying critical data.

Band et al. [35] assessed the usability, defined as ease of use and user satisfaction, and reliability, referring to the consistency and accuracy of results, of various XAI methods in healthcare. Contextual Importance and Utility (CIU) was reported to have higher usability than LIME and SHAP by being more transparent and generating faster explanations in CNNs. The combination of XGBoost with SHAP demonstrated both high accuracy and interpretability, leading to greater acceptance by medical experts, indicating both reliability and usability. One of the studies they assessed indicated that SHAP often presents a balanced performance across usability and reliability metrics compared to LIME and Anchors. To Anchors' advantage, it had the highest average time to output explanations (a usability aspect related to efficiency), while SHAP had the lowest. However, the review highlights the ongoing challenge of establishing a solid framework for evaluating the utility and usability of XAI methods in medical applications.

Rao et al. [36] used a Naive Bayes (NB) classifier for autonomous disease prediction for diabetes, heart disease, and breast cancer datasets. They applied the model-agnostic frameworks LIME and SHAP to interpret the predictions generated by the model. A key finding was the ability of both LIME and SHAP to visually represent the contribution of different features to individual predictions for each disease. For instance, LIME highlighted glucose level as a significant positive factor in diabetes prediction, while SHAP identified it as the highest contributing attribute. Similarly, for heart disease, LIME indicated exercise-induced angina as a risk factor, which was also deemed the most important feature by SHAP. In the breast cancer prediction, both frameworks provided feature importance visualizations for understanding benign and malignant classifications. Their study did not offer much beyond praise of the possibilities LIME and SHAP have in the industry.

Previous evaluations of XAI techniques have largely focused on visualization, intrinsic, and distillation methods. [37] discuss fuzzy logic as a distinct and neglected category of XAI

with the potential to enhance interpretability by addressing uncertainties inherent in medical data. The paper discusses two main ways fuzzy logic can be integrated with AI for enhanced explainability: layer-based and prediction-based approaches. In layer-based, fuzzy logic can be implemented within CNNs to improve interpretability through techniques like fuzzy clustering of features, which provides more clarity into learned patterns. Prediction-based fuzzy logic is used after CNNs produce a prediction. It converts model outputs into interpretable insights. They suggest that embracing fuzzy logic, in addition to established XAI, can contribute to trustworthy, intelligible, and ethical AI in healthcare.

In the financial sector, AI systems rely on principles of explainability, fairness, privacy, accountability, transparency and soundness that defines the base for the AI systems in this field [38]. In this context XAI's main goal is to provide explainability, and interpretability. However, those two concepts cannot easily be achieve concurrently, it depends on the model used in the AI system, for example decision trees and random forest even though they explain the rules they follow, when the nodes in the model grow the model loses interpretability [39]. Therefore, there is a necessity to analyze the complexity of XAI from the interest and needs of the stakeholders. In this regard, SHAP presents a linear model that captures the underlying workings of the AI system that we want to explain, the coefficient is represented as the significance of a particular feature in the model [40]. That makes SHAP a great tool for identifying explanatory variables for instance in Credit Risk assessments [41], and responsibility and accountability in auditing [42].

[43] compares SHAP and GradCAM, for understanding decisions made in human activity recognition (HAR), particularly using graph convolutional networks on skeleton data. The study evaluates the strengths and weaknesses of each method using two real-world datasets, including one for cerebral palsy detection. SHAP offered detailed feature importance, while GradCAM provides spatial visualizations of important body regions. The research concluded that GradCAM is preferable in situations where a quick response is needed such as real-time diagnostics but may miss out on other components which may be sensitive to a successful diagnosis and explanation. SHAP, on the other hand, is effective in detailed explanations of features but requires more computation and time as well as suffers from feature dependencies. They suggest using both GradCAM and SHAP in tandem to give an all-encompassing result.

[44] developed a cardiovascular disease prediction application that evaluated different ML models. They found XGBoost to be the most accurate and utilized it with SHAP for improved transparency and interpretability. They would further explore a combination of SHAP and LIME to provide improved local explanations.

IV. Comparison of Related Works

Ghassemi et al. [45] point out several important issues with the XAI techniques being used today. One of the significant concerns they bring up is the possibility of confirmation bias, which occurs when models produce answers that match users' expectations and might cause them to have unjustified faith in the model's logic. They also highlight a paradox: explanations may make users less aware of potential biases in the model rather than more cautious. Users may become less vigilant because they believe that an explainable model would automatically detect and reduce biases. Another critical limitation they discuss is the lack of performance guarantees in popular XAI techniques like LIME and SHAP. These methods rely heavily on human interpretation, introducing an additional layer of subjectivity and potential error. Essentially, post-hoc explanations add uncertainty rather than clarifying the AI's decision-making process. Overall, they argue that current XAI methods might create a misleading sense of transparency and trust rather than genuinely improving the reliability of AI systems. These explanations simply make users feel more confident without addressing the possible issues within the models themselves. Criticism regarding interpreters of neural networks, specifically [46] investigate adversarial attacks in a heatmap. Their work involved making perturbations in images resulting in large changes in interpretations.

[47] raise the question of ethical concerns and quality assurance with regards to XAI in healthcare. They state the explanations are as important to patients as they are to doctors. The integration of AI methods is largely dependent on doctors, which can only happen if they are correct, accurate, reliable, and trustworthy.

Most of the research aimed at the healthcare sector are enamored with finding AI solutions to facilitate and help practitioners with the vast amount of data they need to process. The two XAI solutions that are researched on the most are LIME and SHAP. They are likely favorites as they are model agnostic, and can work with most data types. Researchers seem to utilize these XAI not for the necessary improvement of the XAI itself, but in their research for developing an improved model that can then be used in healthcare for disease prediction for example.

In the financial sector, XAI models are restricted by policies and laws regarding the data and levels of privacy, in the information managed by the AI model. SHAP, in this regard as data exploration tool, has a major restriction compared to other sectors with less restrictions due to privacy and law constraints.

V. Methodology

This study conducts a systematic survey of literature on SHAP across four major sectors: healthcare, finance, cybersecurity and law. In this way, this study categorizes findings based on sector-specific considerations.

A. Literature Review

1) Research Questions: This study aims to answer the following research questions:

- RQ1: What are the primary strengths and limitations of SHAP across different sectors?

- RQ2: How is SHAP evaluated across sectors, and what metrics are most commonly used (e.g., fidelity, stability, fairness, real-time performance)?
- RQ3: What AI models are most commonly used in conjunction with SHAP?

2) Inclusion and Exclusion Criteria: To ensure the studies selected for review are relevant and of acceptable quality, we applied specific inclusion and exclusion criteria, as summarized in Table I.

3) Search Strategy: The search process involved searching major academic databases:

- IEEE Xplore
- ACM Digital Library
- SpringerLink

4) Search String: To retrieve relevant studies, the following search query was used:

> ("Explainable AI" OR "XAI" OR "Explainable Artificial Intelligence" OR "SHAP") AND ("healthcare" OR "finance" OR "cybersecurity" OR "law") AND ("disease prediction" OR "credit and risk assessment" OR "malware detection" OR "anomaly detection" OR "judicial decision-making" OR "legal risk assessment" OR "survey")

B. Comparative Analysis

1) Data Extraction: For each study, the following information was extracted:

- **Study Information:** Author(s), year of publication, title, journal/conference name.
- **Sector:** Healthcare, finance, cybersecurity, law.
- **Model Type:** Type of AI/ML model used (e.g., CNN, decision tree, random forest).
- **Metrics Evaluated:** Fidelity, trustworthiness, usability, real-time performance, etc.
- **Key Findings:** Key insights about strengths, weaknesses, use case and how SHAP performed.

2) Comparative Evaluation: The extracted data was analyzed and compared based on the following factors:

- **SHAP Use Case By Sector** – Examining how SHAP is used across sectors.
- **AI Model Influence** – Evaluating how AI model choices in different sectors influence the selection of XAI techniques.
- **SHAP Strengths and Weaknesses** – Analyzing reported strengths and weaknesses of SHAP by sector.
- **Comparison of Other Taxonomies** – For survey papers, examining the differences between proposed taxonomies.

C. Taxonomy Development

Based on the literature review and comparative analysis, we propose a sector-based taxonomy that considers specific use cases, AI models, and input data type.

VI. TAXONOMY

Based on the information gathered during the literature review as well as the comparative analysis performed, we propose a sector-based taxonomy that considers the specific use-case and models employed.

TABLE I: Inclusion and Exclusion Criteria

Criteria	Description
Inclusion Criteria	• Studies that focus on SHAP in: – **Healthcare:** SHAP in disease prediction. – **Finance:** SHAP in credit and risk assessment. – **Cybersecurity:** SHAP in malware and anomaly detection. – **Law:** SHAP in judicial decision-making and legal risk assessment. • Studies published in major databases such as **IEEE Xplore, SpringerLink, and ACM Digital Library**. • Studies with a clear and detailed methodology. • Studies published in or translated to English. • Peer-reviewed journal articles and conference papers. • Studies published from **2018–2025** to ensure relevance to current XAI techniques. • Studies that answer at least one research question.
Exclusion Criteria	• Grey literature (e.g., blog posts, white papers). • Studies that do not focus on SHAP applied within the specific sectors and sub-sectors in finance, cybersecurity, law, or healthcare.

TABLE II: Taxonomy for SHAP Across Cybersecurity, Finance, Healthcare and Law

Model	Use Case	Strengths	Limitations	Reference	Publication Year	Data Type
Cybersecurity						
XGBoost	Malicious URL Identification	Efficiency	Complexity, Scalability	[48]	2023	Tabular
XGBoost	Anomaly Detection in Network Traffic	Transparency, Model Refinement	Feature Co-linearity, Scalability	[49]	2022	Tabular
RF	Malware Detection	Robustness, Scalability	Complexity, Overfitting	[50]	2023	Binary
XGBoost	Network Intrusion Detection System	High Accuracy, Enhanced Threat Understanding	Overfitting, Correlation Issues	[51]	2022	Tabular
RFC	Ransomware Detection	High Accuracy	Reliability, trustworthiness	[52]	2021	Tabular
Logistic Regression Model	Intrusion Detection System - APT Detection , Zero Day	Enhanced Interpretability, Complexity resolution	High Computational Cost, Conflicting Explanations	[53]	2022	Tabular
XGBoost and Autoencoder	Network Intrusion	High Detection Accuracy , Enhanced Zero-Day Detection	Overestimation	[54]	2022	Tabular
DNN	Network Intrusion	Improved Transparency , Faster Computation	Computational Cost	[54]	2022	Tabular
Deep Reinforcement Learning	Trust	Malicious AV detection, Higher Trust for Most contributing AVs	Overhead and Time Complexity	[55]	2022	Tabular
Decision Tree	Intrusion Detection	Advanced Anomaly detection	Misclassifications	[56]	2020	Tabular
One vs All , Multi-Class Classifiers	Intrusion Detection	Detailed Attack Characteristics , IDS optimization	Computational Overhead , Dataset Dependency	[57]	2020	Tabular
RF, LR, DT, GNB, SVM	Android Malware Detection System	Feature Contribution Understanding , Transparency	Suboptimal performance for some models	[58]	2022	Tabular
Finance						

Continued on next page

TABLE II (continued)

Model	Use Case	Strengths	Limitations	Reference	Publication Year	Data Type
ANN, XGBoost, SVM and RF	Credit approval	Identifies key features influencing loan		[59]	2021	Tabular
Lasso, CART, RF, XGBoost, MLP	Credit default prediction	Clear ranking among features	Computationally expensive	[60]	2022	Tabular
XGBoost, AdaBoost	Credit scoring	Fairness	Added complexity	[61]	2020	Tabular
CNN	Credit risk assessment	Gives better explanations of local predictions for credit risk assessment than LIME	Global explanations for images are challenging	[62]	2022	Images
XGBoost	Predict funding of loan request	Explainability across models		[63]	2020	Tabular
CNN	Credit Risk assessment	Explainability	Only consider features as a linear combination	[41]	2022	Tabular
RF	Lending to small and medium enterprises	Allows for both local and global explanations		[64]	2022	Numerical
Gradient Boosting Decision Trees	Credit scoring	Model-agnostic		[65]	2021	Tabular
XGBoost	Non-life insurance coverage	Model-agnostic		[66]	2020	Tabular
Healthcare						
LR, RF, DT, NB, XGBoost	Kidney-related disease			[67]	2024	Tabular

Continued on next page

TABLE II (continued)

Model	Use Case	Strengths	Limitations	Reference	Publication Year	Data Type
LR, RF, KNN, XGB, DT, NB, SVM, Improved Explainable Learning-Based Technique (IELBT)	Cardiovascular disease			[68]	2024	Tabular
NB, SVM, voting, XGBoost, AdaBoost, bagging, DT KNN, RF, LR	Cardiovascular disease			[69]	2024	Tabular
DT, KNN, RF, and LR	Mental health	Identifies a subset of inputs which have most utility as features		[70]	2023	Tabular
KNN, rpart (CART), RF	Prostate cancer tissue	Provide importance value to each input feature		[71]	2023	Textual, Tabular
LR, XGBoost	Myocardial infarction	Unpack individual predictions made by XGBoost		[72]	2022	Tabular, Numerical
RF, XGBoost	Thyroid	Quantify the impact of the different features		[73]	2023	Tabular
LR, SVM, RF, XGBoost	Preterm birth	Model independence		[74]	2023	Tabular
Law						

Continued on next page

TABLE II (continued)

Model	Use Case	Strengths	Limitations	Reference	Publication Year	Data Type
XGBoost	Legal decision making	Good plausibility, Pinpoints contribution score of each feature	Inconsistent faithfulness	[29]	2022	Tabular
RF	Predicting inmate misconduct	Clarified how features influenced decision-making	Assumes that features are independent	[27]	2024	Tabular
GBM	Predicting criminal offence in adolescents	Provides both high-level insights and individual predictions		[75]	2024	Tabular
Gradient Boosting Classifier	Classification of final legal decisions		Complexity in large datasets	[76]	2024	Textual
BERT	Legal judgment prediction	Uncertainty detection in model predictions by analyzing descriptive statistics	May not fully capture divergent reasoning when judges disagree	[77]	2024	Tabular

VII. Discussion

This survey analyzes how SHAP is employed across multiple sectors, providing domain-specific insights into its use across healthcare, cybersecurity, law and finance. A key advantage of SHAP is its ability to quantify individual feature importance, making it a preferred interpretability tool across various fields [59]. This benefit is consistently highlighted across all the sectors, suggesting that this is a primary factor in the choice of SHAP, regardless of the domain.

Different sectors prefer specific machine learning models based on their unique challenges and requirements; this indicates that the XAI model should be restricted to a particular domain and level of abstraction. Therefore, the complexity to develop a universal taxonomy that integrates all the areas of knowledge in a single taxonomy. XAI wants to enhance trustworthiness of the stakeholders in the AI model predictions.

In finance, particularly credit and risk assessment, XGBoost emerges as one of the most frequently used models [59]–[61], [63]. In a study by Lusinga et al. [59], four models: Artificial Neural Networks (ANN), XGBoost, Support Vector Machines (SVM), and Random Forest (RF) where compared on the same credit approval dataset. Their findings indicate that XGBoost outperformed the other models, making it a preferred choice for credit approval processes. Most of the studies included tabular data and it was found that SHAP struggled with global explanations for images [62].

Similarly, in law, specifically judicial decision-making and legal risk-assessments, the data used with SHAP was mostly tabular data [27], [29], [75], [77]. Textual data was also used [76] but was mostly converted into tabular data [77]. The same is seen in the cybersecurity sector, particularly malware and anomaly detection, as well as in healthcare. This trend signifies that SHAP works best with tabular data compared to other data formats.

Only papers in the law sector that compared XAI techniques, such as SHAP and LIME [29], included a strategy to evaluate the effectiveness of SHAP. This highlights a gap in the evaluation of XAI techniques and shows the heightened importance of transparency in the legal sector, as emphasized by Górski et al. [28] and Luo et al. [29], compared to the other sectors studied. Similarly, many articles did not explicitly mention the limitations of SHAP. Instead, they used it solely to evaluate and explain their models, without considering the need to assess SHAP's output itself. Among the papers reviewed, Luo et al. [29] was the only one that specified metrics for evaluating XAI techniques, which was based on faithfulness (diffAUC) and plausibility. There has been relatively limited research on SHAP in the legal domain, with most studies emerging only recently in 2024. As more studies are conducted in this field, there is likely to be an increased focus on developing evaluation metrics for SHAP, as well as other XAI techniques, considering the critical need for transparency in order to meet regulatory requirements in the legal sector.

Across all fields, the primary benefit of SHAP highlighted is its ability to show the contribution of each feature to the model's predictions. This is cited as being the main reason for the choice of SHAP compared to other XAI techniques in various studies [29], [59]. Another widely recognized benefit highlighted across studies is SHAP's ability to provide both local and global explanations [76], [78]. On the other hand, a common drawback stated is the computational complexity, especially in large datasets [76], which makes it resource-intensive [48], [50], [53]–[55], [57], [60], [61], [76].

References

[1] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, S. Mack *et al.*, *Principles of neural science*. McGraw-hill New York, 2000, vol. 4.

[2] V. Rajaraman, "Johnmccarthy—father of artificial intelligence," *Resonance*, vol. 19, pp. 198–207, 2014.

[3] R. Aluvalu, M. Mehta, and P. Siarry, "Explainable ai in health informatics," *Springer*, 2024.

[4] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[5] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[6] A. Panesar and A. Panesar, "Ethics of intelligence," *Machine learning and AI for healthcare: big data for improved health outcomes*, pp. 207–254, 2019.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[8] A. Bibal, M. Lognoul, A. de Streel, and B. Frénay, "Legal requirements on explainability in machine learning," *Artificial Intelligence and Law*, vol. 29, no. 2, pp. 149–169, 2021. [Online]. Available: https://doi.org/10.1007/s10506-020-09270-4

[9] I. Benedetto, A. Koudounas, L. Vaiani, E. Pastor, L. Cagliero, F. Tarasconi, and E. Baralis, "Boosting court judgment prediction and explanation using legal entities," *Artificial Intelligence and Law*, pp. 1–36, forthcoming.

[10] L. State, A. B. Colmenarejo, A. Beretta, S. Ruggieri, F. Turini, and S. Law, "The explanation dialogues: an expert focus study to understand requirements towards explanations within the gdpr," 2025. [Online]. Available: https://arxiv.org/abs/2501.05325

[11] B. Brożek, M. Furman, M. Jakubiec, and B. Kucharzyk, "The black box problem revisited. real and imaginary challenges for automated legal decision making," *Artificial Intelligence and Law*, vol. 32, no. 2, pp. 427–440, 2024. [Online]. Available: https://doi.org/10.1007/s10506-023-09356-9

[12] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence applications in cyber security: State-of-the-art in research," *IEEE Access*, vol. 10, pp. 93 104–93 139, 2022.

[13] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (xai) in finance: a systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 8, p. 216, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10854-8

[14] J. Gupta and K. Seeja, "A comparative study and systematic analysis of xai models and their applications in healthcare," *Archives of Computational Methods in Engineering*, 04 2024.

[15] C. A. Zhang, S. Cho, and M. Vasarhelyi, "Explainable artificial intelligence (xai) in auditing," *International Journal of Accounting Information Systems*, vol. 46, p. 100572, 2022, 2021 Research Symposium on Information Integrity Information Systems Assurance. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1467089522000240

[16] D. V. Kute, B. Pradhan, N. Shukla, and A. Alamri, "Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review," *IEEE Access*, vol. 9, pp. 82 300–82 317, 2021.

[17] P. Hacker, R. Krestel, S. Grundmann, and F. Naumann, "Explainable ai under contract and tort law: legal incentives and technical challenges," *Artificial Intelligence and Law*, vol. 28, no. 4, pp. 415–439, 2020. [Online]. Available: https://doi.org/10.1007/s10506-020-09260-6

[18] L. Górski, B. Kuźniacki, M. Almada, K. Tyliński, M. Calvo, P. M. Asnaghi, L. Almada, H. Iñiguez, F. Rubianes, O. Pera, and J. I. Nigrelli, "Exploring explainable ai in the tax domain," *Artificial Intelligence and Law*, 2024. [Online]. Available: https://doi.org/10.1007/s10506-024-09395-w

[19] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, and T. B. Moeslund, "Explainable ai and law: An evidential survey," *Digital Society*, vol. 3, no. 1, p. 1, 2023. [Online]. Available: https://doi.org/10.1007/s44206-023-00081-z

[20] T. Speith, "A review of taxonomies of explainable artificial intelligence (xai) methods," in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2239–2250. [Online]. Available: https://doi.org/10.1145/3531146.3534639

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[24] 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, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[25] E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh, "SHAP-based explanation methods: A review for NLP interpretability," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4593–4603. [Online]. Available: https://aclanthology.org/2022.coling-1.406/

[26] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," 2018. [Online]. Available: https://arxiv.org/abs/1711.11279

[27] F. M. Oliveira, M. S. Balbino, L. E. Zarate, F. Ngo, R. Govindu, A. Agarwal, and C. N. Nobre, "Predicting inmates misconduct using the shap approach," *Artif. Intell. Law*, vol. 32, no. 2, p. 369–395, Mar. 2023. [Online]. Available: https://doi.org/10.1007/s10506-023-09352-z

[28] L. Górski, S. Ramakrishna, and J. M. Nowosielski, "Towards grad-cam based explainability in a legal text processing pipeline. extended version," in *AI Approaches to the Complexity of Legal Systems XI-XII: AICOL International Workshops 2018 and 2020: AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 154–168. [Online]. Available: https://doi.org/10.1007/978-3-030-89811-3_11

[29] C. Luo, R. Bhambhoria, S. Dahan, and X. Zhu, "Evaluating explanation correctness in legal decision making," *Proceedings of the Canadian Conference on Artificial Intelligence*, 05 2022.

[30] A. s. o. Salieh, "A perspective on explainable artificial intelligence methods: Shap and lime," *arXiv preprint arXiv:2305.02012*, 2023. [Online]. Available: https://arxiv.org/abs/2305.02012

[31] Z. Sadeghi, R. Alizadehsani, M. A. CIFCI, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladík, S. Nahavandi, and P. M. Pardalos, "A review of explainable artificial intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, p. 109370, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790624002982

[32] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, p. 4793–4813, Nov. 2021. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2020.3027314

[33] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable ai techniques in healthcare," *Sensors*, vol. 23, no. 2, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/634

[34] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, "From blackbox to explainable ai in healthcare: Existing tools and case studies," *Mobile Information Systems*, vol. 2022, no. 1, p. 8167821, 2022.

[35] S. S Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823001302

[36] S. Rao, S. Mehta, S. Kulkarni, H. Dalvi, N. Katre, and M. Narvekar, "A study of lime and shap model explainers for autonomous disease predictions," in *2022 IEEE Bombay Section Signature Conference (IBSSC)*, 2022, pp. 1–6.

[37] N. Y. Murad, M. H. Hasan, M. H. Azam, N. Yousuf, and J. S. Yalli, "Unraveling the black box: A review of explainable deep learning healthcare techniques," *IEEE Access*, vol. 12, pp. 66 556–66 568, 2024.

[38] C. Maree, J. E. Modal, and C. W. Omlin, "Towards responsible ai for financial transactions," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 16–21.

[39] ——, "Towards responsible ai for financial transactions," in *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2020, pp. 16–21.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[41] C. Cardenas-Ruiz, A. Mendez-Vazquez, and L. M. Ramirez-Solis, "Explainable model of credit risk assessment based on convolutional neural networks," in *Mexican International Conference on Artificial Intelligence*. Springer, 2022, pp. 83–96.

[42] R. Müller, M. Schreyer, T. Sattarov, and D. Borth, "Reshape: Explaining accounting anomalies in financial statement audits by enhancing shapley additive explanations." New York, NY, USA: Association for Computing Machinery, 2022.

[43] F. Tempel, D. Groos, E. A. F. Ihlen, L. Adde, and I. Strümke, "Choose your explanation: A comparison of shap and gradcam in human activity recognition," 2024. [Online]. Available: https://arxiv.org/abs/2412.16003

[44] H. F. El-Sofany, "Predicting heart diseases using machine learning and different data classification techniques," *IEEE Access*, vol. 12, pp. 106 146–106 160, 2024.

[45] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[46] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," 2017.

[47] P. Raif, R. Suchanek-Raif, and E. Tkacz, "Explainable ai (xai) in healthcare - tools and regulations," in *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*, 2023, pp. 151–152.

[48] N. Kaur and L. Gupta, "Enhancing iot security in 6g environment with transparent ai: Leveraging xgboost, shap and lime," in *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*. IEEE, 2024, pp. 180–184.

[49] N. H. A. Mutalib, A. Q. M. Sabri, A. W. A. Wahab, E. R. M. F. Abdullah, and N. AlDahoul, "Explainable deep learning approach for advanced persistent threats (apts) detection in cybersecurity: a review," *Artificial Intelligence Review*, vol. 57, no. 11, p. 297, 2024.

[50] F. S. Prity, M. S. Islam, E. H. Fahim, M. M. Hossain, S. H. Bhuiyan, M. A. Islam, and M. Raquib, "Machine learning-based cyber threat detection: an approach to malware detection and security with explainable ai insights," *Human-Intelligent Systems Integration*, pp. 1–30, 2024.

[51] I. Uysal and U. Kose, "Analysis of network intrusion detection via explainable artificial intelligence: Applications with shap and lime," in *2024 Cyber Awareness and Research Symposium (CARS)*. IEEE, 2024, pp. 1–6.

[52] V. Heydari and K. Nyarko, "Fairness in machine learning for cybersecurity: Enhancing trust through feature importance and shap analysis," in *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2024, pp. 1–6.

[53] M. Naif Alatawi, "Enhancing intrusion detection systems with advanced machine learning techniques: An ensemble and explainable artificial

intelligence (ai) approach," *Security and Privacy*, vol. 8, no. 1, p. e496, 2025.

[54] P. Barnard, N. Marchetti, and L. A. DaSilva, "Robust network intrusion detection through explainable artificial intelligence (xai)," *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.

[55] G. Rjoub, J. Bentahar, and O. A. Wahab, "Explainable ai-based federated deep reinforcement learning for trusted autonomous driving," pp. 318–323, 2022.

[56] R. R. Karn, P. Kudva, H. Huang, S. Suneja, and I. M. Elfadel, "Cryptomining detection in container clouds using system calls and explainable machine learning," *IEEE transactions on parallel and distributed systems*, vol. 32, no. 3, pp. 674–691, 2020.

[57] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73 127–73 141, 2020.

[58] M. M. Alani and A. I. Awad, "Paired: An explainable lightweight android malware detection system," *IEEE Access*, vol. 10, pp. 73 214–73 228, 2022.

[59] M. Lusinga, T. Mokoena, A. Modupe, and V. Mariate, "Investigating statistical and machine learning techniques to improve the credit approval process in developing countries," in *2021 IEEE AFRICON*, 2021, pp. 1–6.

[60] A. A. Robisco and J. M. C. Martínez, "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction," *Financial Innovation*, vol. 8, no. 1, p. 70, 2022. [Online]. Available: https://doi.org/10.1186/s40854-022-00366-1

[61] J. M. Hickey, P. G. D. Stefano, and V. Vasileiou, "Fairness by explicability and adversarial shap learning," 2020. [Online]. Available: https://arxiv.org/abs/2003.05330

[62] C. Cardenas-Ruiz, A. Mendez-Vazquez, and L. M. Ramirez-Solis, "Explainable model of credit risk assessment based on convolutional neural networks," in *Advances in Computational Intelligence*, O. Pichardo Lagunas, J. Martínez-Miranda, and B. Martínez Seis, Eds. Cham: Springer Nature Switzerland, 2022, pp. 83–96.

[63] A. Stevens, P. Deruyck, Z. V. Veldhoven, and J. Vanthienen, "Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 1241–1248.

[64] G. Babaei, P. Giudici, and E. Raffinetti, "Explainable artificial intelligence for crypto asset allocation," *Finance Research Letters*, vol. 47, p. 102941, 05 2022.

[65] W. Liu, H. Fan, and M. Xia, "Step-wise multi-grained augmented gradient boosting decision trees for credit scoring," *Engineering Applications of Artificial Intelligence*, vol. 97, p. 104036, 01 2021.

[66] A. Gramegna and P. Giudici, "Why to buy insurance? an explainable artificial intelligence approach," *Risks*, vol. 8, p. 137, 12 2020.

[67] S. K. Ghosh and A. H. Khandoker, "Investigation on explainable machine learning models to predict chronic kidney diseases," *Scientific Reports*, vol. 14, no. 1, p. 3687, 2024.

[68] P. C. Bizimana, Z. Zhang, A. H. Hounye, M. Asim, M. Hammad, and A. A. A. El-Latif, "Automated heart disease prediction using improved explainable learning-based technique," *Neural Computing and Applications*, vol. 36, no. 26, pp. 16 289–16 318, 2024.

[69] H. F. El-Sofany, "Predicting heart diseases using machine learning and different data classification techniques," *IEEE Access*, vol. 12, pp. 106 146–106 160, 2024.

[70] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *npj Digital Medicine*, vol. 6, no. 1, p. 6, Jan 18 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00751-9

[71] A. Ramírez-Mena, E. Andrés-León, M. J. Alvarez-Cubero, A. Anguita-Ruiz, L. J. Martinez-Gonzalez, and J. Alcala-Fdez, "Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107719, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260723003851

[72] A. Moore and M. Bell, "Xgboost, a novel explainable ai technique, in the prediction of myocardial infarction: A uk biobank cohort study," *Clinical Medicine Insights: Cardiology*, vol. 16, p. 11795468221133611, Nov 8 2022. [Online]. Available: https://doi.org/10.1177/11795468221133611

[73] L. Bellantuono, R. Tommasi, E. Pantaleo, M. Verri, N. Amoroso, P. Crucitti, M. D. Gioacchino, F. Longo, A. Monaco, A. M. Naciu, A. Palermo, C. Taffon, S. Tangaro, A. Crescenzi, A. Sodo, and R. Bellotti, "An explainable artificial intelligence analysis of raman spectra for thyroid cancer diagnosis," *Scientific Reports*, vol. 13, no. 1, p. 16590, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-43856-7

[74] I. K. Kokkinidis, E. Logaras, E. S. Rigas, I. Tsakiridis, T. Dagklis, A. Billis, and P. D. Bamidis, "Towards an explainable ai-based tool to predict preterm birth," *Studies in Health Technology and Informatics*, vol. 302, pp. 571–575, May 18 2023.

[75] J. W. Suh, R. Saunders, E. Simes, H. Delamain, S. Butler, D. Cottrell, A. Kraam, S. Scott, I. M. Goodyer, J. Wason, S. Pilling, and P. Fonagy, "Predicting criminal offence in adolescents who exhibit antisocial behaviour: a machine learning study using data from a large randomised controlled trial of multisystemic therapy," *European Child & Adolescent Psychiatry*, 2024. [Online]. Available: https://doi.org/10.1007/s00787-024-02592-7

[76] O. A. A. Francia, M. N. del Prado, and H. Alatrista-Salas, "Exploring the interpretability of legal terms in tasks of classification of final decisions in administrative procedures," *Quality & Quantity*, vol. 58, no. 5, pp. 4833–4857, 2024. [Online]. Available: https://doi.org/10.1007/s11135-024-01882-1

[77] C. Erdoğanyılmaz, "A new explainable ai approach to legal judgement prediction: Detecting model uncertainty and analyzing the alignment between judges and models," in *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2024, pp. 1–6.

[78] R. Massafra, A. Fanizzi, N. Amoroso, S. Bove, M. C. Comes, D. Pomarico, V. Didonna, S. Diotaiuti, L. Galati, F. Giotta, D. L. Forgia, A. Latorre, A. Lombardi, A. Nardone, M. I. Pastena, C. M. Ressa, L. Rinaldi, P. Tamborra, A. Zito, A. V. Paradiso, R. Bellotti, and V. Lorusso, "Analyzing breast cancer invasive disease event classification through explainable artificial intelligence," *Frontiers in Medicine*, vol. 10, p. 1116354, Feb 2 2023. [Online]. Available: https://doi.org/10.3389/fmed.2023.1116354