Transparency Beyond Accuracy: A Comparative Study of Explainable AI in Credit Scoring and Medical Diagnosis

Hammaad Rizwan
School of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
hammaad.20221729@iit.ac.lk
20221729

Aadhavan Saravanakumar
School of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
arkhash.20221213@iit.ac.lk
20221213

Vinuka Silva
School of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
vinuka.20222185@iit.ac.lk
20222185

Yenuka Rajapaksha
School of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
yenuka.20221359@iit.ac.lk
20221359

***Abstract*—With the unceasing development in machine learning, deep learning and Artificial Intelligence as a whole, the demand for providing reasoning for the decisions and predictions prove to be paramount. This review paper discusses the importance of explainability in using Artificial Intelligence across the domains of credit risk scoring and the medical sector. The primary objective of this review paper is to compare and contrast the necessity of explainability when decisions are made using Artificial Intelligence. These decisions could prove to cause significant effects in these industries. The ethical and regulatory necessities that cause the need for transparency in the domains are rigorously examined. The examination suggests how explainability in credit scoring is driven primarily by factors concerning legal requirements and rationality, whereas the medical sector utilises explainability to augment freedom from suspicion, maintain patient centred care, ethical and moral implications, identifying errors and detecting bias. The findings as a result of the review done suggest on a surface level that while explainable Artificial Intelligence(XAI) benefits both domains, the methodologies and techniques to achieve explainability differ from sector to sector. This research spotlights the importance of context in highlighting how and why AI models should be explainable.**

Keywords—Credit Risk, Bias Detection, Explainable Artificial Intelligence (XAI), Deep Learning and Machine Learning

I. Introduction

The usage of machine learning, deep learning and other artificial intelligence techniques into decision making in the fields of credit scoring and health has led to impressive revolutions. Nevertheless, despite these advancements, the challenge of making decisions derived by these AI models explainable and interpretable is highly strenuous considering the gravity of these domains. These domains have profound consequences on the lives of individuals on a daily basis. Credit scoring plays a major role in the lives of humans as it determines the provision of loans, rentals, housing and sometimes the opportunities of employment. On the other hand, the medical sector is improving rapidly using technology to prevent and cure most harmful diseases. Therefore, applying deep learning algorithms such as object detection to define diseases is a highly sensitive topic. As a result doctors and other experts would need to know the exact steps on how the model makes a final decision. In a recent study titled ‘Explainability of deep neural networks for MRI analysis of brain tumours’ (2022), a program called NeuroXAI was created which provides transparency of how a traditional blackbox deep learning model makes a prediction in classifying brain tumours from MRI images.

As a result these studies highlight the importance of how explainable AI is crucial to domains which include high stakes such as the Finance and Medical sector.

II. Methodological Analysis

A. Credit Risk Scoring

Various methodologies have been used in recent studies aimed at improving the explainability in machine learning and deep learning models used for credit risk scoring. While the complexity of these artificial intelligence models increase, the need for transparency and interpretability has become of utmost importance. This is to ensure the models meet legal and regulatory requirements implemented in the financial industry. This chapter inspects sources of data, techniques for processing the datasets, architectures of the various models, techniques to implement explainability to the results and metrics deployed to validate the explainability provided by the different techniques.

1) Data Sources and Datasets

A variety of datasets from different sources were used to construct effective models for credit scoring. Most datasets used were well known datasets from reliable sources and followed a time series pattern. The diverse range of features proved to be worthy in evaluating the performance of scoring models across different studies.

A German credit dataset [7] was used which contains information on customers from a German bank. The dataset had features related to financial status of the customers, purpose of the loan, employment status and a score to define how worthy the customer is to provide a loan. The dataset had an equal balance of current and delinquent customers.

Another dataset used in one of the research papers is the Australian credit dataset [7] which is commonly used by researchers for benchmarking their models. The dataset has a combination of categorical and numerical features relating to credit history and census information of the customers. The advantage of this dataset is that due its mixed types of features, it could be used to evaluate the explainability of the models across categorical and numerical data.

A home loan equity dataset [7] was used in a study, which mainly aimed at financial histories, loan values and ownership equity of the customers. The dataset contained several features which proved to be useful when evaluating the explainability of the scoring models.

The Credit Card default prediction dataset [17] was a high dimensionality dataset with details of over 30,000 credit card holders in Taiwan. The features of the dataset were related to credit limits, history of customer payments, demographics and billing amounts of the customers over a semiannual period. The industrial level quality of the dataset makes it ideal to evaluate the explainability of the models.

2) *Data Preprocessing*

In most studies, preprocessing raw data was cited as a vital step in building a scoring model. This is because deep learning models require numerical data as input to the model, however financial datasets generally consist of a mix of categorical and numerical features. Thus, below are some of the most noted techniques used to preprocess the raw data.

One of the most commonly used techniques was Weight of Evidence(WOE) Binning [7] which is used in converting categorical variables into numerical ones. This technique is implemented by calculating the likelihood of a default event occurring within a specified category of features. This was used in the German credit dataset where categorical features such as loan purpose and employment status were preprocessed using WOE for each category, resulting in an easier mode of identifying potential delinquent customers by the model.

Another preprocessing method used was One Hot Encoding. This was used on features which could not be ranked easily. This method converts each categorical column by dedicating a separate column to each unique occurrence in the categorical column and converting the values in the new columns to a binary value. For example, the account status feature in the Australian credit dataset contains either Current or Delinquent. This was one hot encoded to be represented as separate binary features.

Information Value(IV) Calculation [7] was another method used to identify and select the most predictive features. The power of each feature in predicting the target variable of the dataset is depicted by the Information Value. Common practices involve removing features with low IV value to reduce noise in the dataset. Furthermore, IV calculation is combined with WOE binning to give priority to features which contribute most to the stability and efficiency of the model while maintaining required levels of explainability.

Moreover, with regards to missing data, they were imputed with statistical measures such as the mean and the mode of the distributions. In some studies, the data was normalised and scaled to ensure the distribution lies within a specified range and contribute equitably to the decision making process. This is essential for deep learning models, as large differences in the distribution of the input features could largely affect the learning process of the models.

Such preprocessing techniques on data ensure the input data for the machine learning and deep learning models are in a suitable and valid format ensuring the model training is efficient while maintaining interpretability.

3) *Model Development*

The explainability of credit risk models are greatly hinged on the construct and structure of the machine learning and deep learning architectures. This review discusses several different deep learning architectures while considering the interpretability and explainability of the models as a focal point.

Ensemble machine learning models have been the most commonly used architectures in most studies related to credit scoring. These include Gradient Boosting Machines (GBMs) [11] and Random Forests [15]. These are combined with deep learning architectures to boost the predictive power and explainability. Ensemble learning works by combining the strengths and weaknesses of multiple models and extracting the optimal performance. This helps improve the overall accuracy and robustness. These ensemble techniques tend to outperform deep learning approaches at times and produce high levels of interpretability, particularly when explainability techniques such as LIME and SHAP are used.

Long Short Term Memory (LSTM) Networks is a deep learning architecture, commonly used in credit scoring when the data involves a time series attribute. LSTMs prove to excel at capturing time series features and trends, which make them effective in scoring the credit risk of an individual over a period of time.

Another novel model architecture and concept which was used involved 2D Convolution Neural Networks(CNNs) in [7]. This involves converting tabular data into images in a format the CNN architecture can comprehend. This is done to extract the image processing power of the CNN. The CNN proves to be effective in processing structured credit data which are binned and one hot encoded into a two dimensional matrix as pixel values. In this concept, each customer's financial data is converted into a 2D matrix [11] and the model trains to classify the image based on the creditworthiness. The Convolution Neural Network functions by detecting patterns in the matrix that depicts the relationships between the financial features and attributes. Pooling layers have been implemented in the architecture of the CNN. Pooling layers are used as a dimensionality reduction technique to focus on the most essential features. These are followed by fully connected layers which are merged to produce a prediction as an output.

Deep Multi- Layer Perceptrons (DMLPs) [11] are another method used in credit risk modelling. They consist of several layers of neurons which are interconnected, where each neuron in the architecture represents a feature which is learned by the model. DLMPs prove to be highly flexible and are commonly used on datasets which contain mixed data types and consist of non-linear relationships among the attributes.

An interesting approach was using textual descriptions of customer transactions and performing transfer learning on a multilingual BERT model [12].

Among the several model architectures were Restricted Boltzmann Machines(RBMs), Deep Belief Networks(DNBs), Autoencoders, Discretized Interpretable Multi-Layer Perceptrons(DIMLPs).

4) *Experimental Setup and Evaluation*

Proper functional testing and diligent experimental setups have to be implemented to ensure the models consist of robustness and reproducibility before being validated for production in the industry. This setup involves training the model, validating, testing and inferencing on new, unseen data.

With regards to the training of the model, the model is trained on a subset of the data while implementing techniques such as gradient descent to optimise the weights for the deep learning model. The validation of the trained model is conducted on the validation set of the dataset, and is used to fine-tune the hyperparameters such as learning rate, batch size and architecture(number of layers/neurons) of the model with regards to deep learning.

Several evaluation metrics have been used to evaluate the models.

1. The most common metric used is accuracy which is the overall percentage of correct predictions made by the model. This provides an overall view of the model performance. This was also measured using the Brier score which is specific for binary classifications [12].
2. Another metric used was the recall(sensitivity) which depicts the model's ability to correctly identify samples which belong to the positive (delinquent) class. This is essential in industrial implementations to identify customers who have a high risk of delinquency to ensure the financial organisation can minimise losses incurred on the organisation itself.
3. Specificity is used to identify the number of negative samples (current customers) classified correctly by the model. This is important to ensure that customers with good credit portfolios are not classified as high risk individuals.
4. The area under the Receiver Operating Characteristic (ROC) curve provides a detailed evaluation of the ability of the model to classify between current and delinquent/ default customers while setting different thresholds for the model. It is a good evaluation metric to assess the performance of the model in terms of both sensitivity and specificity.

5) *Explainability techniques*

Different model explainability techniques have been discussed extensively in these research papers. Explainability refers to the ability to explain the model's behaviour, justify and provide insights into the outputs of the model.

Since deep learning models are known to be black box models, it is difficult to understand the decisions made by these deep neural networks. This occurs due to several reasons. The primary reason is the complexity of the neural networks; these architectures have billions of parameters which consist of non- linear interactions. This makes it tedious to trace the flow from the point of input to the point of producing the predictions. Another reason is the use of non-linear activation functions to transform the outputs produced by one layer as the input of the next layer. Due to their non-linearity, they are difficult to comprehend and trace through the architectural flow of the model.

Despite there being a lack of solid techniques/ tools to interpret the functionality of a deep learning model and provide explainability, certain tools exist which provide insights into the predictions made to a reliable extent.

1. Gradient-weighted Class Activation Mapping (Grad-CAM)[7] was the solitary technique used in providing explanations for the CNN approach of converting each customer's portfolio into a 2D matrix. This technique works by highlighting the sections of input image which contributed the most to the prediction, thus providing insights into how the prediction was made by the model.

2. Local Interpretable Model-agnostic Explanations (LIME) was used as a comparison for the 2D matrix approach. This involved making tweaks in the data and seeing how the predictions of the model change. For example, if the change of a particular pixel in the matrix causes a change in the prediction of the model, the feature pointing to that particular pixel is deemed important. This is followed by building a simple linear model around the modified inputs. This is then used to approximate the behaviour of the CNN model and thereby making the 'black-box' nature of the CNN more comprehensible.

3. SHapley Additive exPlanations (SHAP) [17] were used to calculate the contribution of each feature towards the model prediction. This method sets a value to each feature which depicts the importance of the said feature to the prediction. SHAP proves to be invaluable in providing both local and global explainability.

4. Saliency maps are a type of image which highlights the spatial support of a particular class of an image. This method was deployed mainly while using the CNN approach of using two dimensional matrices.

In addition to this, the effectiveness of the explainability tools is measured by both qualitative and quantitative means. Qualitative analysis involves domain experts such as credit risk analysts and risk officers assessing the explanations provided by these tools such as LIME, Grad-CAM and SHAP. On the other hand, quantitative methods involve metrics such as Feature Importance Agreement, Fidelity and Comprehensibility.

Feature Importance Agreement refers to the evaluation of the importance of features from the explainability methods against known risk factors. If a high agreement value is reached, this proves the explanations are consistent with the domain knowledge. Fidelity refers to how well the explainability tool is able to approximate the predictions of the more complex model, with high fidelity referring to the tool's ability to replicate the model's behaviour. Comprehensibility refers to how straightforward the predictions are to a human user from the respective field(credit risk officer). This is ranked against past user experiences.

These techniques provide an overall understanding of the predictions made by the model and provide support to confirm the model provides enough explainability to be trusted in making decisions at an industrial level with very minimal to no fault.

6) Limitations of Current Methodologies

Despite the availability and existence of techniques mentioned above, there prove to be certain limitations when it comes to setting these models in a production scenario in the industry.

Discussed below are some of the shortcomings of the currently existing methods of improving explainability in credit risk scoring.

1. Primarily, most of the explainability techniques are dependent on the architecture of the model. For example, the approach of converting the profile of each customer from tabular format to a two dimensional matrix before being fed into a CNN may be limited only to this particular model architecture. This may not be able to generalise on other models such as Recurrent Neural Networks (RNN) and Transformers. This approach may simply work for models designed only for image processing.

2. The training and inferencing process of deep neural networks requires high levels of computational resources. This requirement augments when these models are integrated with explainability techniques such as SHAP which requires the computation of the Shapley values for each attribute being fed to the model. In a production scenario, the requirement of such resources may prove to be unprofitable in terms of manpower and cost [17].

3. Moreover, results obtained from libraries such as LIME and SHAP need to be rigorously evaluated with domain experts to ensure they align with known values [7]. This is of utmost importance as the results may not align with the real world domain knowledge which could be catastrophic as it involves the financial status of individuals- the primary factor affecting the lifestyle of humans. Moreover, more complex representations of model explainability such as saliency maps may prove to be difficult to comprehend for domain experts without the required training, which makes such methods inapplicable in real life scenarios.

4. Bias in training data is one of the primary concerns in modelling credit risk. If there is a bias in the training data possibly due to imbalance in demographic factors and other constituents, this

could cause the model to sustain these biases. Explainability techniques could highlight which features contribute influentially in the model making a decision, but they are not able to detect existing biases in the data.

7) *Future Directions*

While there has been significant improvements in the field of using Deep Learning for credit risk modelling and prediction, there lies an incessant need for improvement. Some of the key points to be researched on in future are as follows;

1. Though several different types of machine learning and deep learning architectures have been used in research study, novel approaches involving different model architectures such as Graph Neural Networks (GNNs) and Recurrent Neural Networks (RNNs) could be tried. Though CNNs have shown promising results when processing tabular data into image format, it may not be effective for all types of credit data. Transformers have proven to be successful in Natural Language Processing (NLP), they could possibly be effective in modelling credit scoring by converting the tabular data into sequences of textual data [12].
2. More focus has to be given into feature engineering in identifying underlying behaviours. Though explainability techniques provide insights into the predictions of the model, they depend on the underlying features. New features could be created to reflect on complex relationships between attributes and patterns in time series features. These features could boost the accuracy and interpretability of scoring models.
3. Fairness is one of the factors of utmost importance. Future research should be done on how explainability techniques could be used not only to understand the predictions made by the model but also to reduce bias and ensure fairness [15]. A suggestion might be amalgamating these explainability techniques with algorithms which are aware of fairness to ensure decisions are not made based on race, gender or social/ economic status.
4. Another point to be considered is the integration of domain experts with these studies [15]. Though there exist several experts in the field of credit scoring, there lies a lack of these experts who have a sound knowledge of Artificial Intelligence and most importantly the explainability techniques. This proves to be a pitfall as these individuals may not be able to interpret these tools. Thus novel tools have to be created which could be understood by domain experts who have a minimal technical knowledge.

B. *Medical Diagnosis*

Different methodologies have been used in current studies aimed at enhancing the explainability in machine learning and deep learning models used for the medical sector. As the complexity of these artificial intelligence models increase, the need for transparency and interpretability has become of utmost importance. This is to ensure that the decisions made by the models make sense and are understandable for the clinicians. This will help the healthcare experts make reliable and informative decisions with confidence when diagnosing patients. This chapter inspects sources of data, architectures of the various models, techniques to implement explainability to the results and metrics deployed to validate the explainability provided by the different techniques.

1) *Data sources and Datasets*

The papers mainly focus on the explainable AI types used in the medical field rather than specifically giving details about the datasets used, however there are few data sources which were mentioned.

Reference [20] mentions that Electronic Health Records (EHRs) are used as a source of data for healthcare predictive modelling.Its used as they consist of patient medical histories which are maintained by healthcare providers. There are certain difficulties in using this data, such as combining data from different providers and large amounts of missing data due to patients' irregular visits. MIMIC is an example of an open-source EHR. Scientific literature such as PubMed and MEDLIN are also important data sources which could be used as they consist of reliable findings. Natural Language processing and text mining techniques are used to extract the relevant information from these scientific sources. For medical images, the MedPix database by the National Library of Medicine could be utilised. Genomic data is provided by the National Human Genome research. It consists of resources which give diseases associated with certain genetic variations. Epidemiological data is predominantly used to assess nutritional and health status of a population (EX - National Health and Nutrition Examination Survey (NHANES)).

Reference [8] discusses the use of Radiology Picture Archiving and Communication Systems (PACS) to extract 320,000 clinically important lesions from the human body. Reference [3] uses the MedNIST dataset which is used to mainly classify CT scans and MRI images into several categories like AbdomenCT and BreastMRI.

Reference [16] uses a heart disease dataset which consists of preexisting data related to heart strokes. Most of the papers discussed focus on the explainable AI techniques rather than on the datasets used to train the model, due to this reason there is limited information on the datasets used.

2) Model Development

The main focus of the papers were to explain and give a detailed description of the explainable AI techniques. They do not mention how the models were selected and implemented in detail.

3) Experimental Setup and Evaluation

Reference [13] mainly discusses the conceptual aspects of explainable AI and causability in the field of healthcare. While it doesn't mention a quantitative or specific experimental setup, the paper gives some qualitative insights as to how the explainability can be assessed. As an example for human post-hoc explanation, qualitative insights were obtained by a pathologist for liver biopsy diagnosis. This example was utilised by the authors to illustrate a human centric approach to explain decisions by the machine learning models by comparing the pathologist explanations with the AI system output. The pathologist also provided features to evaluate (macroscopical and microscopical) liver pathology. This was considered to be an ante-hoc explanation.

Reference [19] gives a general framework to evaluate explainability of AI systems. Using application grounded evaluation was one recommended technique (domain experts test the explanation). Human-grounded evaluation and functionally grounded evaluation were also recommended in this paper as other possible techniques to evaluate explainability.

Reference [16] makes use of explainable AI to assess the accuracy of different types of classification models which were used to predict a heart score risk. Quantitative metrics were not provided as to how the explainable AI were used to assess the model accuracies.

Reference [3] conducts an analysis in PubMed for the explainable AI techniques used in medical imaging.

4) Explainability techniques

Many researches primarily focus on reviewing and categorizing XAI methods and their applications rather than providing detailed experimental results and comparisons. Therefore, a comprehensive analysis of performance and explainability based on specific experimental results from the sources is limited. However, some researches offer some insights into the performance and explainability of certain XAI methods through specific examples and analyses.

1. Saliency Maps
   Reference [19] states that the visualization techniques used in saliency methods have helped researchers identify flawed reasoning in classification problems, improving the understanding and debugging of AI models. The authors discuss the use of Grad-CAM to visualise pleural effusion in radiographs and the application of CAM for interpretability in brain tumour grading, indicating their potential in medical image analysis.
   The study [3] compares different saliency methods, including DeepLIFT, LRP, and guided backpropagation, for classifying insomnia using physiological network data. It revealed variations in the attribution maps generated by each method. Additionally, found LRP and guided backpropagation to be most effective in generating coherent attribution maps for Alzheimer's disease classification.
2. LIME
   Reference [19] describes LIME as a method that explains a model's decision by highlighting the importance of input features. For example, LIME identified specific symptoms as crucial for predicting flu in a patient.
   Reference [13] explains that LIME balances local fidelity, ensuring the explanation accurately reflects the model's behaviour locally, while minimising the complexity of the explanation for better interpretability. [presents a study utilising LIME for Alzheimer's disease detection, where it pinpointed specific brain regions as crucial for patient classification.
3. SHAP
   The research [13] suggests that CIU (Contextual Importance and Utility), surpasses LIME and SHAP in usability for decision-making support, as it provides better transparency and faster explanation generation. It presents a comparative analysis of LIME, SHAP, and Anchors for explaining tabular and text data. The results demonstrated that SHAP exhibits a balanced performance in terms of usability and reliability metrics for both data types.
4. Inpainting-Based Occlusion
   Reference [1] proposes an inpainting-based occlusion method (IBO) for evaluating XAI methods. IBO replaces occluded regions with contextually relevant information through inpainting, allowing a more precise assessment of the impact of occluded features on model predictions.
   The study also describes an experiment using the CAMELYON16 dataset to evaluate IBO. The dataset contains 400 whole-slide images (WSIs) of sentinel lymph nodes, with 270 slides having

precise pixel-level annotations. These annotations are used to calculate the Intersection over Union (IoU) metric to compare the heatmaps generated by XAI methods with the ground truth.

5. Occlusion Sensitivity
Reference [8] mentions that occlusion sensitivity is a perturbation based technique that visualises the significance of certain locations of an image for a specific task (ex - classification). It is mainly used for generating visual explanations for deep learning models. This technique perturbs the input image (ex - occluding parts of the image to observe how the prediction would change. This technique is applied in various medical image analysis (ex - differentiating between images of healthy patients and patients with certain conditions, localising lesions and grading disease severity). Certain parameters (size and shape of the occluded area) should be defined, these can influence the results.

6. Class Activation Mapping(CAM)
Reference [8] suggests the use of CAMs to get visual explanations for convolutional neural networks(CNN). It is specifically designed for CNNs with a certain architecture where fully connected layers at the end of the network are replaced by global average pooling(GAP) applied to the last convolutional feature maps. CAM highlights the areas of the input image that are influential in the CNN's decision for a certain class. Use cases of CAM in medical imaging are localization of diabetic retinopathy, analysis of histology images, analysis of brain MRI, analysis of chest X-rays, analysis of fundus photography, analysis of endoscopy images and analysis of dermatoscopy.

7. Feature importance
Reference [8] states that feature importance is a method of post hoc explanation. The purpose of feature importance is to provide insight of learned relationships by analysing trained neural networks. Global feature important values can provide insights into how much features affect the output across the whole dataset.

8. Randomised Input Sampling for Explanation(RISE) of Black-box Models
[3] defines RISE as a perturbation based method which uses random occlusion patterns to recognize regions of the image that contribute most to the model's output. These random occlusion patterns are made by sampling small binary masks (ex - 7 x 7 pixels) and interpolating small masks to larger resolutions. Subsampling permits for the recognition of the important areas of the image. This technique is also limited to the dependance of pre-defined parameters such as the number of epochs, number of masks created and the kernel size. The biggest challenge is to get the most suitable values for the above mentioned parameters to get a balanced accuracy and efficiency.

9. Causability
Reference [16] proposes the use of causability as a metric to assess the standard of the explanations provided by the explainable AI systems to the interested parties. In simple terms it's the measure of how well a human expert in the medical field can understand the explanation provided by the XAI system.

10. Integrated Gradients
Reference [3] mentions the use of integrated gradients, which is a technique used to give an accurate and complete attribution for each of the features used. This is done by calculating the output's average gradient with respect to the input.

11. Layer-Wise Relevance Propagation (LRP)
Reference [19] discusses the use of LRP XAI technique as it explains the predictions provided from deep neural networks (DNN). This is done by decomposing the output given by the DNN in terms of the relevance of each input feature. Starting from the output layer it propagates the relevance scores backwards through the DNN.

5) Limitations of Current Methodologies

Although there's a wide range of explainable AI techniques which could be utilised in the healthcare field as discussed above , there are a significant number of limitations as well. Discussed below are some of the main limitations.

1. Lack of Specificity
Many of the current XAI techniques fail to give the exact features that are directly responsible for a model's prediction. These methods could highlight areas which are not significant as well which could lead to misinterpretation by the users.[8]

2. Dependance on Visual Explanations
In certain use cases visual explanations might not be enough to give a complete and meaningful understanding of how the model came up with a certain prediction [8].

3. Depending on Predefined Parameters
When utilising techniques like LIME, the user should provide a certain set of predefined parameters. These could significantly change the

results which could make it difficult to reproduce. Selecting optimal parameters upfront is a difficult task which requires expertise [3].

4. Sensitivity to Architectural Variations and Implementation
   Some of the techniques could be more reliable with models of a certain architecture whereas they could be less reliable with a certain group of models. The proper selection of the most suitable XAI technique is important [8]
5. Challenge in Quantitative Evaluation
   Measuring the quality of the explanations provided by the techniques is a major challenge as it's mainly done by healthcare experts, which could lead to biases and inconsistency as it depends on the person. [13]
6. Lack of Guidelines and Best Practices
   The lack of a predefined set of specific guidelines for the evaluation and implementation of XAI techniques is a major concern [8].

6) Future Directions

Even though there were a significant number of improvements with explainability in AI systems, there are many paths that need to be researched in order to bridge the gap between the AI systems and human understanding of how a certain AI system came up with a prediction.

1. Developing Medical Imaging specific XAI
   This could be done by integrating domain specific knowledge bases into the models, which would provide clinically relevant information. The XAI systems will be able to learn the anatomical structures and characteristics related to diseases relevant to specific medical images. This would lead to more detailed and specific explanations [3]
2. Beyond Visual Explanations
   Providing textual explanations using natural language description of the model's reasoning would provide a detailed and specific explanation. This would be helpful when the visual representation itself is insufficient. Using example based explanations is also a good method to research as it will make the explanations more simple and understandable [16].
3. Standardised Evaluation of XAI methods
   Developing quantitative metrics to measure the effectiveness of XAI explanations is critical for comparing existing methods [8].
4. Establishing Guidelines for XAI in the field of medicine
   A set of guidelines should be developed to validate and implement explainable AI in the field of healthcare, this could increase transparency and the responsible use of AI in this field [8]

III. Benchmarking

XAI Method	**Type**	**Application Area**	**Pro and Con**	**Evaluation Metric**
Saliency Maps	Post-hoc visual	Medical Imaging (CT, MRI)	Pro: Intuitive heatmap visualisation Con:Limited to visual data	Accuracy, AUD, Fidelity
Layer-wise Relevance Propagation (LRP)	Post-hoc Explanation	Histopathology, Medical Imaging	Pixel-wise relevance, precise Con:High computational cost	Fidelity, Trust
Class Activation Mapping (CAM)	Post-hoc Visual	Radiology, Medical Imaging	Highlights relevant regions Con:Less generalizable for non-visual data	Sensitivity, Precision
Gradient-weighted CAM (Grad-CAM)	Post-hoc Visual	Tumour Detection, Breast Cancer	Robust feature mapping Con:Limited to specific model	Sensitivity, Specificity
SHAP (Shapley Additive Explanations)	Model-agnostic	Predictive Modelling in Healthcare	Clear feature importance ranking Con: Computationally expensive	Consistency, Accuracy
LIME (Local Interpretable Model-agnostic Explanations)	Post-hoc Explanation	Diagnosis, Surgery Prediction	Local explanations model-agnostic	Interpretability, Fidelity
Inpainting-based Occlusion (IBO)	Post-hoc Visual	Histopathology Image Processing	Reduces OoD artifacts, High computational cost	LPIPS, AUC
Deconvolution	Backpropagation-based	Imaging (CT, MRI)	Reveals feature layer patterns, Sensitive to model structure	Relevance, Fidelity

IV. Summary of Results

TABLE I. Experimental Results Summary

Paper	Evaluation Metric/ XAI Technique	Score
Credit Scoring		
Making deep learning-based predictions for credit scoring explainable	**Grad-CAM** Accuracy AUC Brier Score	91% 0.87 0.09
	LIME Accuracy AUC Brier Score	74.% 0.64 0.26
	Saliency Maps Accuracy AUC Brier Score	76% 0.64 0.26
Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction	**CreditNetXAI** Accuracy Sensitivity specificity	83.5% 0.8823 0.9879
Medical		
Application of Explainable AI in Medical Health	Validation Accuracy	98.58%
	Testing Accuracy (Biassed Datasets)	94%
	Testing Accuracy (Unbiased Datasets)	86%
Applications of Explainable Artificial Intelligence in Diagnosis and Surgery	Accuracy (Allergy Diagnosis - kNN, SVM, C 5.0, MLP, AdaBag, RF)	86.39%
	Sensitivity (Allergy Diagnosis - kNN, SVM, C 5.0, MLP, AdaBag, RF)	75%
	Accuracy (Rule-Based Breast Cancer Diagnosis)	60.81%
From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies	Accuracy of classification models with LIME: Bayesian Rule Lists Multilayer Perceptron Dempster-Shafer Classifier RNN Gradient Descent	75.6% 76.4% 61.2% 66.9% 83.8%

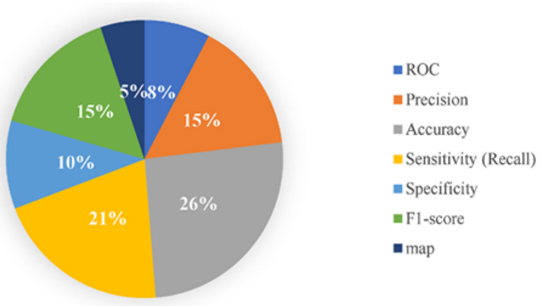

Fig. 1. The share of various metrics to implement DL model's performance. [2]

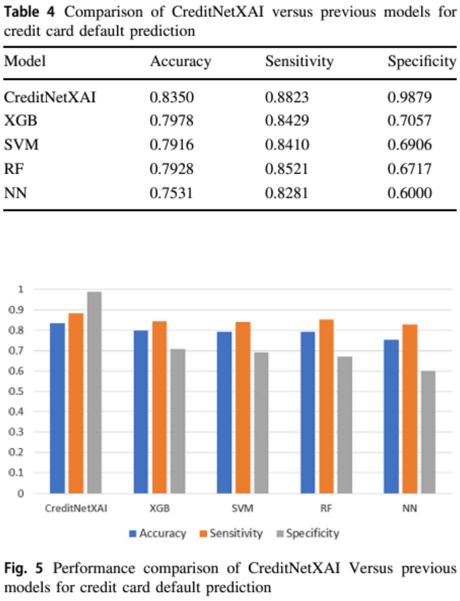

Table 4 Comparison of CreditNetXAI versus previous models for credit card default prediction

Model	Accuracy	Sensitivity	Specificity
CreditNetXAI	0.8350	0.8823	0.9879
XGB	0.7978	0.8429	0.7057
SVM	0.7916	0.8410	0.6906
RF	0.7928	0.8521	0.6717
NN	0.7531	0.8281	0.6000

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CreditNetXAI
XGB
SVM
RF
NN
Accuracy
Sensitivity
Specificity

Fig. 5 Performance comparison of CreditNetXAI Versus previous models for credit card default prediction

Fig. 2. Performance Comparison of CreditNetXAI [17]

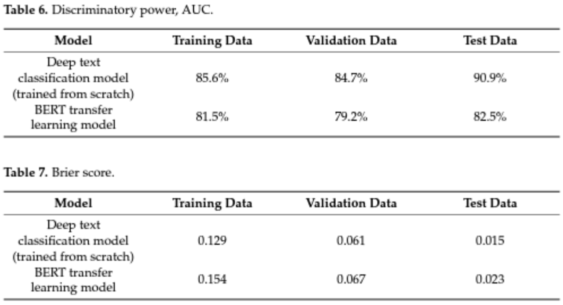

Table 6. Discriminatory power, AUC.

Model	Training Data	Validation Data	Test Data
Deep text classification model (trained from scratch)	85.6%	84.7%	90.9%
BERT transfer learning model	81.5%	79.2%	82.5%

Table 7. Brier score.

Model	Training Data	Validation Data	Test Data
Deep text classification model (trained from scratch)	0.129	0.061	0.015
BERT transfer learning model	0.154	0.067	0.023

Fig. 3. Evaluation of Deep Text model [12]

VI . Research gap

A. *Technological Gap*

1. Despite the existence of explainability techniques such as Grad-CAM and SHAP, there tend to be limitations in adapting these tools. In the credit scoring sector, transparency is required for complex algorithms whereas in the medical sector, the provided explainability should be interpretable by domain experts. This gives rise to the gap of lack of robust and user-friendly XAI tools which are customizable.

2. Another concern is the trade off between accuracy and explainability. AI models with high accuracy are often less interpretable. In the medical sector, despite the need for explainability, the accuracy of predictions cannot be compromised. This is very similar in the credit sector where explainability is essential for meeting legal requirements.

B. Domain Gap

1. Despite advancements in explainable AI (XAI), there remains a significant research gap which is the absence of general XAI methods throughout a wide range of areas including financial, medical, etc. Current approaches are domain-bound and address particular forms of data and decisions, specific organisational regulations and rules, which makes their applicability and adaptability rather questionable. It is crucial to use a combined method as an interaction between data science and ethics, psychology, and domain-specific knowledge for the creation of more adaptable XAI for different kinds of requests in different sectors. XAI techniques alone, without such a cross-domain, adaptable framework, the valuable tools cannot achieve widespread practical application. Filling this void is important for the development of XAI systems that provide stable, effective and comprehensible solutions that are acceptable across various disciplines in order to benefit society.
2. One of the most understudied area in XAI is the lack of attention to causability that is, not only asking 'how' the AI reached a conclusion, but 'why' it did so, across domains such as healthcare, finance and etc. Modern approaches to XAI focus mainly on interpretability and fail to provide causal explanations needed in professional domains particularly when it comes to understanding AI decision making. Causability needs interdisciplinary work where causal inference, domain specialisation, and ethics are applied in building systems that offer context-sensitive explanations based on causality.. The bridging of this gap is crucial for building AI that generates the right type of understanding and improving their competency in several fields.

VI . Conclusion

'Black box model' is a term that is given to an algorithm that is not transparent about its steps to generate the final prediction. Therefore the bridge in between integrating Deep learning to solve real world problems would be reduced through Explainable AI.

As discussed in various case studies specifically to the credit risk domain we saw that methods such as Grad-CAM, Saliency Maps, SHAP and LIME provide an intuitive way to show the steps taken to come up with the solution. Through these steps the model becomes more transparent thus allowing more firms to integrate deep learning into their workspace, making Explainable AI more prominent.

Causability is linked to the understanding of 'why' things take place in AI Systems which is fundamental to XAI in the medical domain as it narrows the gap between human comprehension and transparency. Users understand how AI systems operate as well as reasons behind its actions and what it can potentially do, by focusing on models and human-centric metrics. This understanding helps in building trust and enhances human AI cooperation and ensures that AI is applied in healthcare in a safe and effective manner.

As a result we see that Explainable AI is important in both of these domains as the outcomes of the model is highly crucial hence we should not only look into enabling XAI in these two domains but for all applications including AI.

References

[1] Afshar, P., Hashembeiki, S., Khani, P., Fatemizadeh, E. and Rohban, M.H., 2024. Ibo: Inpainting-based occlusion to enhance explainable artificial intelligence evaluation in histopathology. arXiv preprint arXiv:2408.16395.

[2] Band, S.S., Yarahmadi, A., Hsu, C.C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A.T. and Liang, H.W., 2023. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, p.101286.

[3] Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M. and Nensa, F., 2023. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. European journal of radiology, 162, p.110786.

[4] Bracke, P., Datta, A., Jung, C. and Sen, S., 2019. Machine learning explainability in finance: an application to default risk analysis.

[5] Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J., 2020. Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 3, p.26.

[6] Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J., 2021. Explainable machine learning in credit risk management. Computational Economics, 57(1), pp.203-216.

[7] Dastile, X. and Celik, T., 2021. Making deep learning-based predictions for credit scoring explainable. IEEE Access, 9, pp.50426-50440.

[8] de Vries, B.M., Zwezerijnen, G.J., Burchell, G.L., van Velden, F.H., Menke-van der Houven van Oordt, C.W. and Boellaard, R., 2023. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Frontiers in medicine, 10, p.1180773.

[9] Fahner, G., 2018. Developing transparent credit risk scorecards more effectively: An explainable artificial intelligence approach. Data Anal, 2018, p.17.

[10] Gramegna, A. and Giudici, P., 2021. SHAP and LIME: an evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, p.752558.

[11] Hayashi, Y., 2022. Emerging trends in deep learning for credit scoring: A review. Electronics, 11(19), p.3181.

[12] Hjelkrem, L.O. and Lange, P.E.D., 2023. Explaining deep learning models for credit scoring with SHAP: A case study using Open Banking Data. Journal of Risk and Financial Management, 16(4), p.221.

[13] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), p.e1312.

[14] Misheva, B.H., Osterrieder, J., Hirsa, A., Kulkarni, O. and Lin, S.F., 2021. Explainable AI in credit risk management. arXiv preprint arXiv:2103.00949.

[15] Nallakaruppan, M.K., Balusamy, B., Shri, M.L., Malathi, V. and Bhattacharyya, S., 2024. An Explainable AI framework for credit evaluation and analysis. Applied Soft Computing, 153, p.111307.

[16] Srinivasu, P.N., Sandhya, N., Jhaveri, R.H. and Raut, R., 2022. From blackbox to explainable AI in healthcare: existing tools and case studies. Mobile Information Systems, 2022(1), p.8167821.

[17] Talaat, F.M., Aljadani, A., Badawy, M. and Elhosseini, M., 2024. Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Computing and Applications, 36(9), pp.4847-4865

[18] Tjoa, E. and Guan, C., 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), pp.4793-4813.

[19] Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G. and Viergever, M.A., 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 79, p.102470.

[20] Yang, C.C., 2022. Explainable artificial intelligence for predictive modelling in healthcare. Journal of healthcare informatics research, 6(2), pp.228-239.

[21] Zhang, Y., Weng, Y. and Lund, J., 2022. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics, 12(2), p.23