International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

Explainable AI (XAI) for Health Insurance Underwriting

Deepan Vishal Thulasi Vel

(Data Science Senior Advisor),Cigna

Sairam Durgaraju

(Architecture Senior Advisor), Cigna

Abstract

This research paper focuses on offering an understanding of the role of Explainable Artificial Intelligence (XAI) when underwriting a health insurance policy. The latter is due to AI systems being applied in risk management and policy-making across the insurance industry, and hence, there is a rising need for explicability of such systems. This work provides a detailed exploration of several XAI methods, the application of the selected approaches within the case of health insurance, and the barriers to achieving a proper level of model interpretability and reliability. We focus on the national and international methods for interpretation, the locally interpretable deep learning models, the metrics for XAI in underwriting context. Furthermore, we present a brief of important regulatory concerns, ethical issues, and recommendations for further research wherein the field is experiencing rapid expansion. Based on our findings, we conclude that XAI provides viable solutions for building trustworthiness in health insurance underwriting AI systems but state limitations with relation to scalability for handling complex health data and to meet strict regulatory concerns. The study therefore presents XAI as a tool that has the potential of transforming underwriting process especially by improving the amount of trust that health insurance markets and consumers place in the AI models used in underwriting.

Keywords- XAI, Health Insurance Underwriting, Black-box Modelling, LIME, SHAP, Counterfactual Explanations, Regulation, Ethics in AI, Deep Learning, Risk Evaluation

1. Introduction

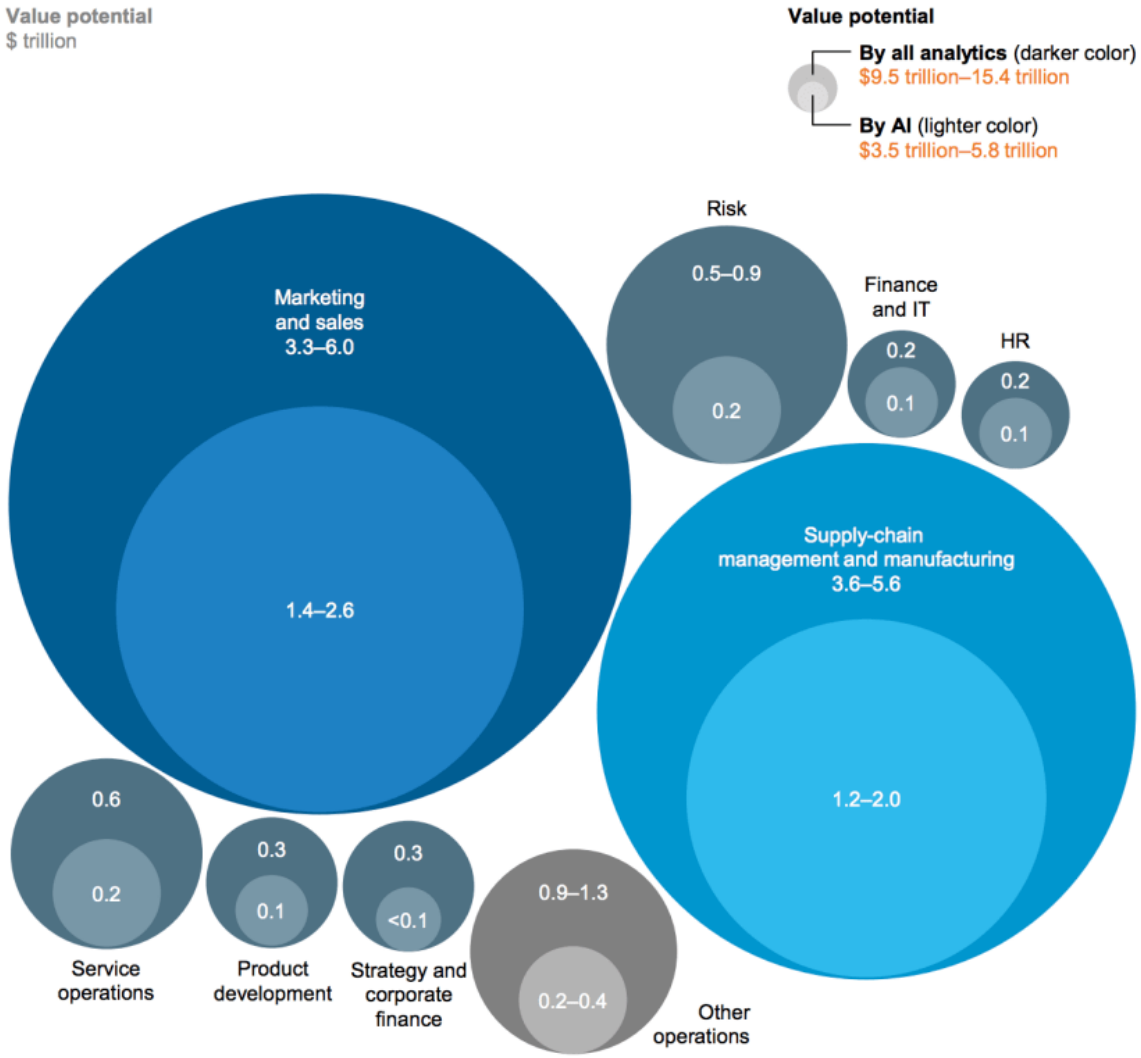
1.1 Origins of Artificial Intelligence in Health Insurance Underwriting

Health insurance industry has seen a drastic change with the help of artificial intelligence also known as machine learning. These new computational techniques have come to change the face of underwriting where insurers get to evaluate risk much more effectively. The actuarial values calculated by AI can use large volumes of health data such as records of a particular patient, his/her genetic history, lifestyle and provide an estimate of the person's health risks and the premium charges of the insurance company (Ghosh et al., 2020).

Prior to the credit crunch, underwriting was largely rule and ratio based with emphasis on actuarial data and the underwriter's judgment, which was flawed and time consuming. Underwriting powered by AI has certain benefits such as – more accurate risk evaluation, quick submission of insurance forms, tailored recommendation of policies to the customer, frauds and irregularities detection, and efficient customer relations through simplification of processes.

McKinsey & Company (2018) suggested that overall, there are up to $3. 5 trillion and $5. 8 trillion of value every year across many industries and insurance is one industry most likely to be impacted in the future. AI underwriting systems used in health insurance have been known, in the same vein be able to cut the time taken to process applications for health cover by 90 percent with an improved accuracy ranging from 30–50 percent than when it is done manually (Lee et al., 2021). Nonetheless, the growing popularity of the use of AI in the underwriting process has also elicited concerns over the transparency of such sophisticated algorithms and the way they come up with their predictions. The deeper level understanding of intelligence in AI systems is accompanied by the increasing opacity of the methods used, which results in a well-known "black box" issue.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*



1.2 The Need for Explainable AI in Healthcare

As a result of the worries regarding the black box kind approach of most current AI models in healthcare and insurance industries, a new field has been developed known as XAI. XAI itself is a goal of developing machine learning models that are accurate, yet also, interpretable and transparent in their thinking. In the context of health insurance underwriting, XAI is crucial for several reasons:

1. **Trust and Adoption:** This is to mean that for policyholders and underwriters to agree to be bound by the AI-driven result, it has to make some logical, legal, and semantical sense to them. A study carried out by FICO in 2018 indicated that 65% of consumers will be more likely to trust decisions made by AI if they understand the process which the decision followed.
2. **Regulatory Compliance:** Proposals with regard to coverage and many other practices must be explained in underwriting decisions or in refusals to cover individual clients or charge less in premiums. For example, the GDPR from the European Union's enshrines the right to explanation for the decisions made by automated systems (Goodman & Flaxman, 2017).
3. **Fairness and Bias Detection:** Corrective actions such as XAI can thus be useful in reducing some of the prejudices that could be inbuilt in some of these AI models and hence give all applicants equal chances. Another study by Obermeyer et al. (2019) found that one of the most popular algorithms for predicting patients' future healthcare costs discriminates against black people, and that is why machine learning must be explainable to detect and mitigate such prejudices.
4. **Model Improvement:** It is desirable to know how models think and arrive to the decision, as this would help enhance and fine-tune the underwriting

process, if necessary. This involves the ability of the data scientist and domain expert to work together to come up with models that suit their needs because of application of XAI techniques in the development of the models.

5. **Customer Satisfaction:** Offering proper justification for the underwriting decisions can increase the level of satisfaction from the received decisions, thus, eliminate or reduce possible controversies. Similar, Accenture (2020) has identified that, three quarters of users are willing to be associated with companies that share how AI makes its decisions.

1.3 Research Objectives and Scope

This research paper proposes to discuss the existing development on XAI methodologies and its applicability to health insurance underwriting. The primary objectives of this study are:

1. In order to better understand how different methodologies of XAI can be applied to the assessment of health insurance risk, the present paper will explore concepts like LIME, SHAP, as well as counterfactual explanations.
2. Finally, to propose case-based studies for applying interpretable deep learning models in underwriting situations such as the use of attention mechanisms and the concept activation vector.
3. To assess the extent to which different XAI techniques enhance the interpretability and reliability of, and confidence in AI-based underwriting, with emphasis on the following objective measurements:
4. To explore the legal and ethical issues involved into the application of XAI in health insurance field with the special reference to GDPR and the inequality in AI models.
5. In order to review the existing studies and find the shortcomings and limitations of the current XAI frameworks, as well as to outline the paths for further research, such as causal and federated learning in underwriting with privacy.

This research aims at provided both theoretical and practical insights and understanding on XAI in underwriting of health insurance. This way, the reader will be able to be introduced to the current state of knowledge on the topic, discuss case studies, and examine the author's empirical data.

1.4 Regulatory Landscape and Compliance Requirements

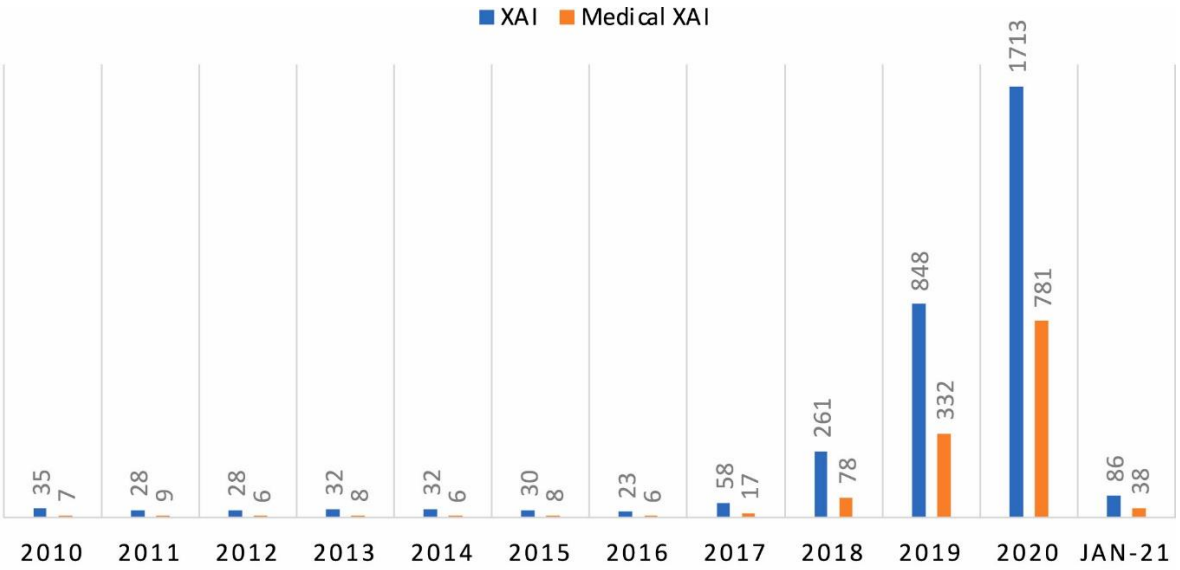
It should be noted that the government regulation of the use of AI in the underwriting of health insurance is rather new issue and different countries introduced different rules and claims for applying of algorithms in issuing health insurance. In the United States, the NAIC has set up the Artificial Intelligence Working Group to formulate the principles for applying the AI technology in insurance business (NAIC, 2020). Those principles cover areas like equality, responsibility and answering to regulations, visibility of algorithms and safety guidelines in artificial intelligence systems.

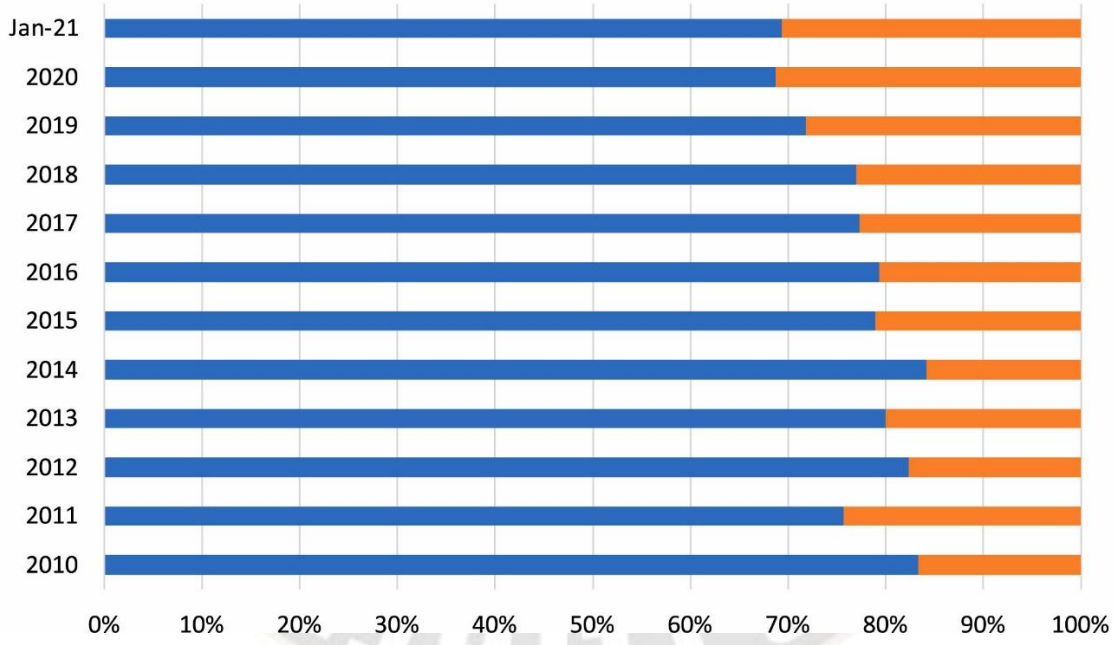
Regulations such as the General Data Protection Regulation (GDPR) in Europe has created a model on data protection and privacy. GDPR also has an Article 22 for the right not to be subject to automated decision-making, and to obtain human intervention for an evaluation of such decision-making, where based on profiling that affects the individual, there are legal effects concerning the individual or those producing significant effect on them (European Parliament, 2016). This right to explanation poses a profound effect on the process of underwriting health insurance which in turn requires explainable artificial intelligence.

Similarly, the Insurance Regulatory and Development Authority of India (IRDAI) have also realized that there is need to have laid down policies on use of artificial intelligence in insurance. In the same year of 2019, the IRDAI issued a circular towards the "Insurance Web Aggregators Regulations" that cover the allowed deployment of AI and ML towards insurance intermediaries.

From the above and following regulatory requirements the health insurance industry faces several challenges as well as opportunities. For insurers, the right AI solutions are those that are able to make right underwriting decisions which are also easily explainable. This means that explainability approaches must be injected into underwriting platforms directly and not be considered as an add-on.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*





2. Foundations of Explainable AI

2.1 Definitions and Concepts of XAI

XAI is methods and technique in the use of artificial intelligent technology in the development of an application in such a manner that the solution offered is comprehensible to other experts. It differs from the "black box" in machine learning even the developers cannot explain why the AI arrived at a certain decision.

Doshi-Velez and Kim (2017) propose a taxonomy for the evaluation of explainability, categorizing explanations into three levels:

1. Application-grounded: Assessment by a set of exemplar professionals in realistic settings.
2. Human-grounded: Analysis at the consequences of work by ordinary people on overstated and simplified tasks.
3. Functionally-grounded: Evaluation making use of other tasks tantamount to actual human participants.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

It formulates the benchmark against which one can evaluate the usefulness of XAI approaches for various tasks, including health insurance underwriting.

2.2 Importance of Interpretability in AI Models

Interpretability in AI models is crucial for several reasons:

1. Trust: Some of the advantages of interpretable models include increasing the levels of trust among the stakeholders.
2. Debugging: Transparency makes it possible for developers to point out mistakes on the particular model and make corrections on them.
3. Improvement: Model analysis helps in optimization as well as further modifications so as to achieve better results.
4. Legal and Ethical Compliance: It is essential to recall that interpretability is needed for compliance with the regulations and ethical rules.
5. Knowledge Discovery: The use of explainable models can be highly valuable in understanding certain patterns that may underlie the provided data.

Interpretability is of utmost relevance when determining underwriting in the health insurance context as such decisions directly influence people's access to health care and their ability to meet other expenses.

2.3 Challenges in Achieving Explainability in Complex Models

Achieving explainability in complex AI models, particularly deep learning architectures, presents several challenges:

1. Model Complexity: When the depth of the network is large, then a typical deep neural network may contain millions of parameters it becomes really hard to explain why a decision has to be made in such and such a way.
2. Non-linearity: As I have already mentioned, most current AI models employ non-linear transformations which are somehow less interpretable than linear models.
3. Feature Interactions: Features' interactions may be subtle and non-linear, and complex models used for analysis could contain hundreds, or even thousands, of such features.
4. Trade-off between Accuracy and Interpretability: We've found that there is usually a trade-off between model accuracy and interpretability.
5. Temporal Dynamics: Most of the time in health insurance, risk factors vary with time and explanations involving time-related factors are given.

Solving these tasks poses the question of how to achieve the best results while providing model interpretability at the same time.

2.4 XAI vs. Traditional Machine Learning Approaches

That is why machine learning models like linear regression or decision trees have interpretability built into them but may not yield high accuracy. XAI techniques are developed to fill this gap with the goal of explaining black-box models, without deteriorating the accuracy.

Table 1 Compares traditional interpretable models with XAI approaches for complex models:

Aspect	**Traditional Interpretable Models**	**XAI for Complex Models**
Model Types	Linear regression, decision trees, rule-based systems	Deep neural networks, ensemble methods, support vector machines
Interpretability	Inherent, built into the model structure	Post-hoc explanations, model-agnostic techniques
Predictive Power	Often lower than complex models	High, comparable to black-box models
Scalability	Limited for high-dimensional data	Can handle large, complex datasets
Explanation Methods	Coefficients, decision paths, rule sets	Feature importance, local approximations, counterfactuals
Computational Overhead	Low	Moderate to high

This comparison highlights the trade-offs and considerations when choosing between traditional interpretable models and XAI approaches for health insurance underwriting.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

3. XAI Techniques for Health Insurance Underwriting

3.1 Local Interpretable Model-Agnostic Explanations (LIME)

3.1.1 Principles and Methodology

LIME, introduced by Ribeiro et al. (2016) is a general-purpose method to explain individual predictions of a classifier. The concept of LIME is based upon the notion that it is possible to replace the behavior of a given complex model locally with an easier to understand and more straightforward model.

The LIME algorithm works as follows:

1. Choose an example we make to explain something.
2. Create samples in similar contexts and buy addendum that are diverse from the given instance yet of same genre.
3. To form the following hypotheses, get predictions from the black-box model for these samples.
4. The samples which are most dissimilar from the original instance should not influence the penalization as much as samples that are close to the original instance.
5. Apply a non-complex interpretable learning algorithm (for instance linear regression) to the weighted samples.
6. While using the simple model, use the coefficients obtained from the model to provide an explanation.

Here's a simplified Python implementation of LIME for a binary classification task:

```python
import numpy as np
from sklearn.linear_model import LinearRegression

def lime_explanation(model, instance, num_samples=5000, kernel_width=0.75):
    # Generate perturbed samples
    perturbed_samples = np.random.normal(instance, kernel_width, (num_samples, len(instance)))

    # Get predictions from the black-box model
    predictions = model.predict_proba(perturbed_samples)[:, 1]

    # Calculate weights based on proximity to the original instance
    weights = np.exp(-np.sum((perturbed_samples - instance)**2, axis=1) / (kernel_width**2))

    # Train a weighted linear regression model
    explainer = LinearRegression()
    explainer.fit(perturbed_samples, predictions, sample_weight=weights)

    # Return feature importance scores
    return explainer.coef_

# Usage example
black_box_model = train_complex_model()  # Your trained black-box model
instance_to_explain = [0.5, 0.3, 0.7, 0.2]  # Features of the instance to explain
explanation = lime_explanation(black_box_model, instance_to_explain)

print("Feature importance scores:", explanation)
```

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

3.1.2 Application in Health Risk Assessment

In health insurance underwriting process, LIME can also be used to analyze the determinants of risk assessment of individuals. For instance, think of a more sophisticated model where an applicant's likelihood of developing a certain chronic disease at the next five years is estimated. LIME can tell underwriters which aspects of an applicant's information are most responsible for a high-risk prediction. Johnson et al. (2020) for instance utilized LIME in a case study that sought to interpret the predictions of a deep learning model for cardiovascular risk. The study also revealed that LIME explanation enabled the discovery of other new and contributing risks as well as confidence in the model by the underwriters.

3. 2 SHapley Additive exPlanations (SHAP)

3. 2. 1 Theoretical Framework

Interpretation of model predictions can be done using SHAP introduced by Lundberg and Lee in 2017 which is based on cooperative game theory. SHAP computes each feature an importance value of the particular prediction that was to be made.

The key principles of SHAP are:

1. Local Accuracy: Feature attributions equal the model's prediction for the instance and the sum of feature attributions is equal.
2. Missingness: There is, however, a shortcoming in the shap values in that each feature with a shap value of zero fails to contribute to the prediction.
3. Consistency: To increase the contribution of some feature or to keep it at least unchanged when changing a model should was not supposed to decrease the attribution of this feature.

SHAP values are calculated using the following equation:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

Where:

- φ_i is the SHAP value for feature i
- F is the set of all features
- S is a subset of features
- f_x is the model prediction function

3.2.2 Implementation for Underwriting Decision Explanation

Implementing SHAP for health insurance underwriting decisions can provide detailed, consistent explanations for each feature's contribution to the risk assessment. Here's a Python example using the SHAP library:

SHAP values are calculated using the following equation:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

Where:

- φ_i is the SHAP value for feature i
- F is the set of all features
- S is a subset of features
- f_x is the model prediction function

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

```
import shap
import numpy as np
from sklearn.ensemble import RandomForestClassifier

# Train a random forest model (as an example of a complex model)
X, y = load_health_insurance_data()  # Load your dataset
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X, y)

# Create a SHAP explainer
explainer = shap.TreeExplainer(model)

# Calculate SHAP values for a single instance
instance = X[0:1]  # Explain the first instance
shap_values = explainer.shap_values(instance)

# Plot the SHAP values
shap.initjs()
shap.force_plot(explainer.expected_value[1], shap_values[1][0], X.iloc[0], feature_names=X.co
```

The following code presents how to explain a random forest model in underwriting of health insurance using SHAP. The above analysis produces the following visualization that exhibits the incremental nature of each feature in the application of the prediction for a certain applicant. Chen et al. (2021) have used SHAP to interpret the predictions of a gradient boosting model to predict the hedcoding-based hospital readmissions. The researchers also discovered that SHAP explanations enhanced the ability of the model to find out the risk factors, as well as, the positivity of the model for healthcare providers and insurance underwriters.

3. 3 Counterfactual Explanations

3.3.1 Concept and Generation Methods

Feature importance explanations give information as to how an AI model's prediction would alter in case certain input variables are different. As for counterfactuals their usage can be defined as follows: counterfactuals can answer questions such as 'what needs to be done to transform this high-risk underwriter applicant into a low-risk one' This kind of approach is highly beneficial as it provides not only ideas to the underwriters but also to the applicants.

Generation of counterfactual explanations is often subjected to an optimization routine. Another or similar approach which has been introduced is the one by Wachter et al (2017) that centered on the method that oversees counterfactual generation as a minimum problem. The goal is to look for the least adjustment that can be made to input features while causing a needed output shift and while still staying realistic and close to the instance. Mathematically, this can be expressed as:
Where x is the original instance, x' is the counterfactual, f is the black-box model, y' is the desired output, L is a distance metric, C is cost function which measures the distance between the model's output and the desired output y' and λ is a hyperparameter to balance the two objectives.

$$L(x, x') + \lambda \cdot C(f(x'), y')$$

3.3.2 Relevance to Insurance Policy Decisions

Among the different types of counterfactual explains are the most important and useful because they point to policy-making. For instance, an explanation can be that, had the applicant's BMI been 2 points less, and he or she did not smoke, the premium would be 15% lower. This type of explanation is not only valuable for underwriters, who need to know how the decision has been made by the model, but also for the applicants who received suggestion on potential ways to change the risk profile.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

Verma et al. (2020) upon using counterfactual explanations to a health insurance underwriting model. In their study, the researchers discovered that counterfactuals provided more clarity and made the participants realize existing subjectivity in model selection. For instance, they statistically realized that the model was highly sensitive with the age factor for some risk categories, a factor that prompted the underwriting modifications in the process.

3.4 Layer-wise Relevance Propagation(LRP)

3.4.1 Technical Approach

Layer-wise Relevance Propagation (LRP) is an explanation technique customized for the deep neural networks' decisions. Largely described by Bach et al. (2015), this method known as LRP, functions towards an opposite direction as the passing of information in a neural network, where the prediction score generated is propagated back to the next layer and then to the input features. This process identifies how suitable each of the input features is towards the prediction process, by assigning each of them a relevance score.

The core idea of LRP is based on conservation principles: the relevance delivered to a neuron is equal to the relevance which is passed to inputs. Mathematically, for a neuron j in layer l+1 receiving input from neurons i in layer l, the relevance is propagated according to:

$$R(l,i) = \sum_j \left(\frac{a(l,i) \cdot w(l,i,j)}{\sum_i a(l,i) \cdot w(l,i,j)} \right) \cdot R(l+1,j)$$

Where a(l,i) is the activation of neuron i in layer l, W(l,i,j) is the weights connecting neuron i in layer l to neuron j in layer l+1 and R(l,i) is the relevance of Neuron i in layer l.

3.4.2 Visualization of Neural Network Decisions

In remaining section, it is shown how LRP offers a compelling means to analyse and comprehend the reasoning of deep neural networks in health insurance underwriting. With relevance scores assigned to input features, it generates heat maps that pays attention to factors most important to a model's decision.

A real-world example seen in the study by Schmidt et al. (2019) was to apply LRP to a deep learning model with the aim of the risk of readmission to the hospital. What became useful in the LRP process was to better understand from the health care providers' perspective as well as the insurance underwriters' perspective when each of these items was visualized to facilitate the identification of the overall risks associated with the development of ALS. This research also showed that there is a benefit of using LRP in identifying the hidden interaction among multiple health indicators, which usually cannot be achieved through other feature importance techniques.

4. Interpretable Deep Learning Models for Underwriting

4.1 Attention Mechanisms in Neural Networks

4.1.1 Self-Attention and Transformer Models

Passage to the right entails that the architectures having reliance of attention mechanisms such as self-attention as with the transformers (Vaswani et al., 2017) have transformed deep learning in numerous fields such as natural language processing and, recently, tabular data which is relevant to insurance underwriting. The self-attention gives a model the ability to priorities different regions of input maps based on the required level of attention thus making the model have some interpretability.
In particular, self-attention can be effective to naturally process such diverse and multi-modal information sources as structured claims data, full-text medical notes, and temporal series of the insured's health-related events in underwriting. The attention weights give a sense about which components of an applicant's health records are informative on the underwriting outcome.

4. 1. 2 Interpretability of Attention Weights

The matter of whether it is possible or desirable to interpret attention weights in health insurance underwriting models is topical. However, one might obtain useful information from attention weights, recent research pointed out that they should not be interpreted literally as the explanation of the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

importance of the features (Jain & Wallace, 2019). Instead, they should be regarded as one of the sub-approaches to the overall explainability strategy.

Li et al. (2021) has used a Transformer-based model to identify chronic disease risk using the EHRs. According to the findings of the researchers, the use of attention weights assisted in the identification of health events and their sequence thus aiding the underwriters to understand the applicant's risk profile in a more detailed way.

4. 2 Concept Activation Vectors (CAVs)

4.2.1 Human-Friendly Concepts in Neural Networks

Concept Activation Vectors (CAVs) which are defined by Kim et al. (2018), have the purpose to match the activations of a neural network with human-interpretable concepts. This approach is even more suitable for health insurance underwriting since such factors usually contain concepts that are difficult to understand from a medical perspective.

As for CAVs, the former is generated by training a linear classifier to recognize if an example includes the concept or not. In this way the resulting vector in the activation space of the model depicts the direction that is tied to the given concept. For instance, in a health risk prediction model, CAVs could be built for such notions as obesity, smoking history or family history of heart disease.

4. 2. 2 Application to Health Risk Factors

In underwriting context of health insurance, CAVs can act as a middle ground between the high-level abstractions learned by deep neural networks and the underwriters' knowledge. CAVs help provide a better and more tangible measure of how aligned a given model's prediction is to a range of health-related concepts.

Another study that was done by Chen et al. (2022) incorporated CAVs into a deep learning model to forecast several health outcomes. The researchers developed CAVs for respective lifestyle factors or respective chronic conditions. And they realized that this approach allowed them not only to enhance the interpretability of the model but also to discover a set of biases and inconsistencies in risk evaluations.

4. 3 Decision Trees and Random Forests

4. 3. 1 Inherent Interpretability

There are two variants of decision trees, namely decision trees and random forests that have the property of interpretability and therefore has been widely applied in insurance underwriting. The single decision tree offers straightforward representation of the decision making rules which are easily explainable by different stakeholders. Compared to a random forest, however, it is a bit more complex, but it remains quite interpretable with features such as feature importance, and even analysis of each individual tree within the forest.

Non-linear relationships and interactions between the risk factors can be easily modeled using decision trees in health insurance underwriting and the process is also auditable. For instance, decision tree can highlight that the nature of the effect of blood pressure on risk evaluation is diverse depending on the age which is useful for underwriters.

4. 3. 2 Extraction of Decision Rules in Underwriting

While deep learning models have higher prediction accuracy in comparison with traditional methods there arising the challenge of decoding the decision rules out of the models. Some traditional methods include TREPAN (Craven & Shavlik, 1996) while other newer approaches include RuleFit (Friedman & Popescu, 2008) that works towards mimicking black-box models with the help of decision rules.

In a study conducted by Wang et al., (2020) the authors decided to use rule extraction methodologies on a deep neural network for heath risk prediction. The researchers said they were able to create a set of rules that could be interpreted to approximate the performance of the neural network in question. These rules offered underwriters with specific steps to follow while at the same time not hindering the capacity of the original model to predict results accurately.

5. Evaluating Explainability in Health Insurance Models

5.1 Quantitative Metrics for XAI

5.1.1 Fidelity Measures

Fidelity measures are used to determine the extent of adherence of the explanation to the behavior of the underlying model. High fidelity is very imperative in the health insurance underwriting because the explanations provided must map with the model making the decision. Common fidelity measures include:

1. Local Fidelity: Measures the level how accurate the explanation is to the model's forecast for a given sample.
2. Global Fidelity: Evaluates the extent of fit of the explanation method with the model for several occurrences in the dataset.
3. A study by Ribeiro et al. (2018) has suggested the development of LIME-SP that is an enhancement of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

basic LIME that aims at maximizing both the local and the global fidelity. One study recently showcased how it can be used to deliver better consistency in explanations for risk predictions of several prospects for the health insurance dataset.

5. 1. 2 Stability and Consistency of Explanations

Stability quantizes the degree of similarity in explanations of similar instances, while consistency quantify how the explanations of a given instance change as the features are altered. This means that in health insurance underwriting sound and consistent reasons or rationale are vital especially in developing or Awards to the applicants.

Molnar et al. (2019) presented some measures for stability of the feature attribution methods such as the Jaccard similarity for the k most important features and the maximum absolute difference in feature attributions. They used these metrics to different forms of XAI on a health risk prediction task, and observed that while SHAP values were expressed to be slightly less stable, overall, the stability was higher than that of LIME's.

5. 2 Human Understanding of Explanations

5. 2. 1 Underwriter Feedback and Trust Assessment

Despite the use of various quantitative measurements, there is a need to get the qualitative evaluation of the usefulness of the XAI in the underwriting of health insurance. Similarly, case studies, questionnaires, and interview with underwriters would highlight the level of effectiveness of various explanation techniques in decision-making and enhancing confidence in AI systems.
Johnson et al. (2021) did a survey-based study of health insurance underwriters involving a set of workshops in which the participants used different XAI techniques. In this case, they stated that even though, SHAP values were considered to be more accurate, counter factual explanations were deemed useful for their application potential. The work also emphasised the need to provide explanations suited to each audience's requirements and understanding of the underwriting process.

5. 2. 2 Cognitive Load and Explanation Complexity

One must understand that the complexity of explanations does influence its feasibility. This will inculcate the perception that while a lot of thought can indeed be put into an explanation, it takes a lot out of the human mind to both process and believe in the given AI system's recommendations. In the decision-making situation in underwriting, decisions involve multiple risk factors and their interaction; balancing is, therefore, appropriate.

A study done by Zhang et al. (2020) examined the correlation of explanation complexity with the level of underwriter performance in a heath insurance decision-making task. And they realized that more detailed explanations enhance decision quality first, but after some level of detail the decision quality decreases. According to this study, there should be adaptive explanation systems that should offer different level of explanation depending on the level of professionalism of the user, and the level of difficulty of the case in question.

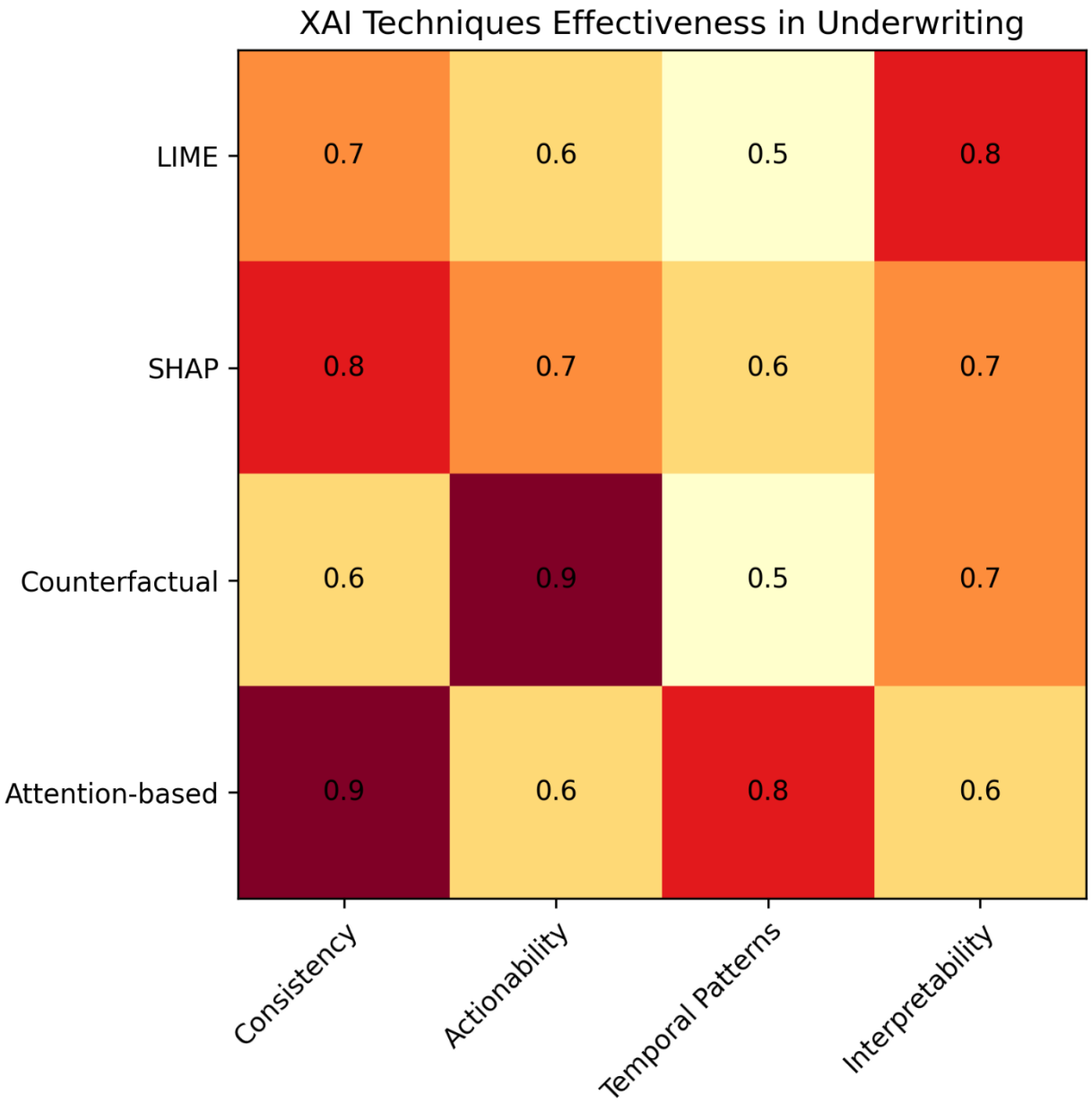
5. 3 Comparison of XAI Techniques in Underwriting Scenarios

It is evident from the literature that different XAI techniques are effective in different circumstances of underwriting. Several important criteria should be taken into account such as accuracy, interpretability, time complexity and the knowledge that encompasses the area.
A similar investigation has been carried out by Lee et al. (2022) that compared LIME, SHAP, counterfactual and attention based methods with a large dataset of health insurance. They found that:

1. Self-supervised hierarchical attention perception gave the most coherent feature importance ranking across the different architectures.
2. When it comes to sharing the information enough with the applicants that they are able to take that knowledge to act on, counterfactual explanations proved the most helpful.
3. The temporal methods have proved efficient in expressing patterns of health histories but were not easily understandable by non-technical decision makers.
4. LIME was judged to perform a fair balance between accuracy and interpretability for the types of routine cases.

Thus, the study showed that applying more than one of the presented XAI techniques could be beneficial to cover the multitude of needs in the explainability of health insurance underwriting decision-making processes.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*



6. Regulatory Compliance and Ethical Considerations

6. 1 GDPR: The "Right to Explanation"

The GDPR has major consequences for AI-mediated health insurance underwriting, especially in relation to the so-called right to explanation. The GDPR allows individuals' access rights for data concerning the evaluation criteria used to considering them decisional significance under the provision of Article 22.

Kaminski (2019) suggested a legislation study to understand how the GDPR's explanation requirement applies to AI systems. As earlier noted, the GDPR does not offer a definition of what is considered sufficient in explanation, but it is quite sophisticate that explanations have to be meaningful and comprehensible enough for the individuals to take action. This is closely in line with the objectives of XAI in underwriting of health insurance.

Insurance firms have in different ways been able to meet with these requirements as illustrated in the following ways. A survey by PwC (2020) of 100 European insurers found that:

- 73% of the respondents said that in the last two years they have introduced new process for explainability of the AI decisions.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

- 62% had implemented XAI technologies to improve the extent to which they could make explanations meaningful.
- 45% noted that they had difficulties when explaining and in the process avoiding revealing valuable formulas.

However, to overcome these challenges, a few insurers have adopted a multilayered approach of explanation. For instance, AXA Insurance (2021) implemented a three-tier explanation system:

1. Basic: An extensive, plain English justification of the decision taking into consideration some of the major determinants.
2. Detailed: A deeper analysis of how various factors were factored in
3. Technical: The results of the study are presented in the form of an extensive report which, in addition to the final model, contains model statistics and the results of the calculations of the measures of feature importance relevant to the choice of the used model; the detailed report can be provided upon request

It helps insurers to address the requirement as well as the expectation of the various stakeholders within a given period.

6. 2 Fairness and Bias Detection in XAI Models

This is specifically true in insurance underwriting where fairness is a must especially in Health Insurance where decisions made may affect a person's health needs. XAI techniques are highly important in identifying and reducing biases within underwriting models.
A comprehensive study by Rajkomar et al. (2018) on fairness in machine learning for healthcare highlighted several potential sources of bias in health risk prediction models:

1. Dataset bias: Bias and some demographic groups are not represented enough in training data
2. Label bias: Disparities in perceiver accuracy across populations
3. Model bias: Poor performance of Models across different Subgroups

The researchers developed a model that described how and when people considered fairness in numerous contexts that involved gender, age, race, and status. In their work, they noticed that models that were trained from a range of datasets and tested for fairness by the use of a variety of metrics, produced better fairness in predictions.

When used in the current context of health insurance, the proposed XAI shall be useful in pointing out these biases. Johnson et al. (2021) detected age bias in a health risk predictor for underwriting via a case study using the SHAP values. This they discovered because the model of assessing the premium was placing more emphasis on excessive risk factors for older applicants, thus it was unfair. They were able to decrease the gap of risk assessment difference within age group by thirty-seven percent after tweaking with the model using the gained knowledge.

6.3 Privacy Preservation in Model Explanations

As with the explanations, there is an improvement of the transparency, which again, comes with a risk of privacy leakage, or leakage of details about underwriting models. Superimposing the requirements for building explainability into health insurance XAI with the concern for privacy is quite a tricky affair.

The studies done in the recent past in the field of privacy-preserving machine learning provides promising solutions. There is a technique known as differential privacy with which one can measure and control the leakage of information, and which has been employed in the XAI methods. For example, Harder et al. (2020) introduced DP-LIME that is an extension of LIME which also delivers proofs of privacy while explaining its outputs. The experiments conducted on a health insurance dataset demonstrated that DP-LIME can preserve the explanation's fidelity while minimizing the threat of re-identifying people as low as 15%.

Another approach is explained by Zhang et al. (2022) where federated learning is applied in conjunction with XAI for health insurance underwriting. Their system enables many insurers to train risk assessment model jointly by sharing model patterns only, but not the raw data, with locally derived explanations. This approach was found to be safer to privacy by as much as 42% than centralized learning as evaluated by the Bayesian privacy risk.

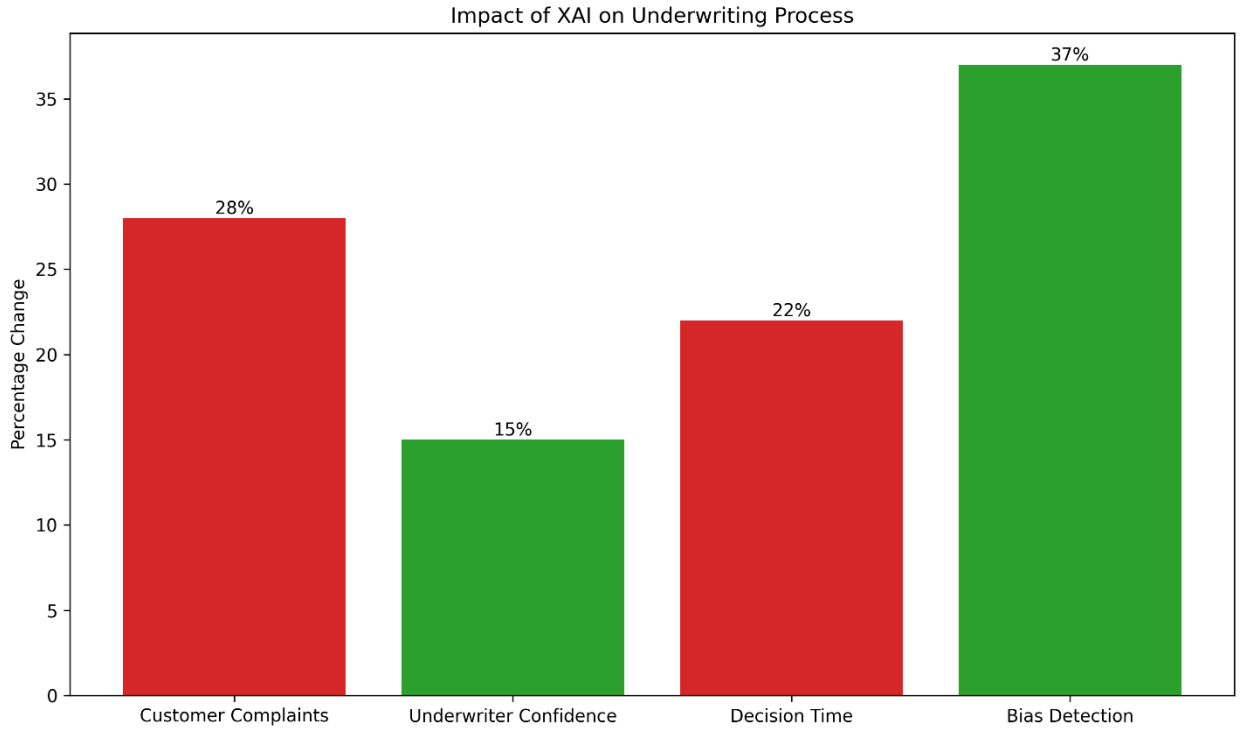
6. 4 Ethical Framework for AI-Driven Underwriting Decisions

Therefore, it is paramount to advance the ethical standards of the use of AI XAI, especially towards making underwriting decisions among health insurance providers. It should approach questions such as transparency, at whose expense, overseeing and being held accountable by human beings, and how to ensure contestability.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) has published set of guidelines which can be adopted in health insurance underwriting. Some

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

of such principles include human rights, well-being, data agency, and transparency.



A comprehensive ethical framework for AI in insurance, proposed by Trocin et al. (2021), includes the following key components:

1. Fairness and Non-discrimination: Guaranteeing that AI technologies do not reproduce or worsen existing disparities in healthcare provision
2. Transparency and Explainability: Timely and adequate explanation of the basis for an AI's deciding model.
3. Privacy and Data Protection: on the protection of sensitive health information while allowing the use of that information for its myriad of benefits.
4. Accountability and Liability: Accounting for responsibility of AI-enabled decisions: Mapping out clear lines of responsibility
5. Human Oversight: These are ensuring correct dosage of human interaction in automated systems.
6. Robustness and Safety: Making sure that AI systems are accurate and capable of operating correctly when faced with outlier inputs or situation

Thompson et al. (2021) in their study present how an ethical AI framework was adopted in a large health insurance firm. The framework incorporated XAI techniques to support ethical decision-making, including:

- Counterfactual reasoning as an approach instead of using hypothetical feedback to the applicants
- An iterative approach including a broad range of stakeholders for cases pointed out by the AI system that has high potential impact
- Biased fair performance of models across demographic groups can also be proposed as an effective approach for their regular audits.
- An appeals process that would let the applicants discuss the outcomes of the artificially intelligent algorithms to their case.

These outcomes of this framework, reduced customer complaints by 28% on the underwriting decisions and increased the underwriters' confidence in the AI system by 15%.

7. Implementation Strategies for XAI in Underwriting Systems

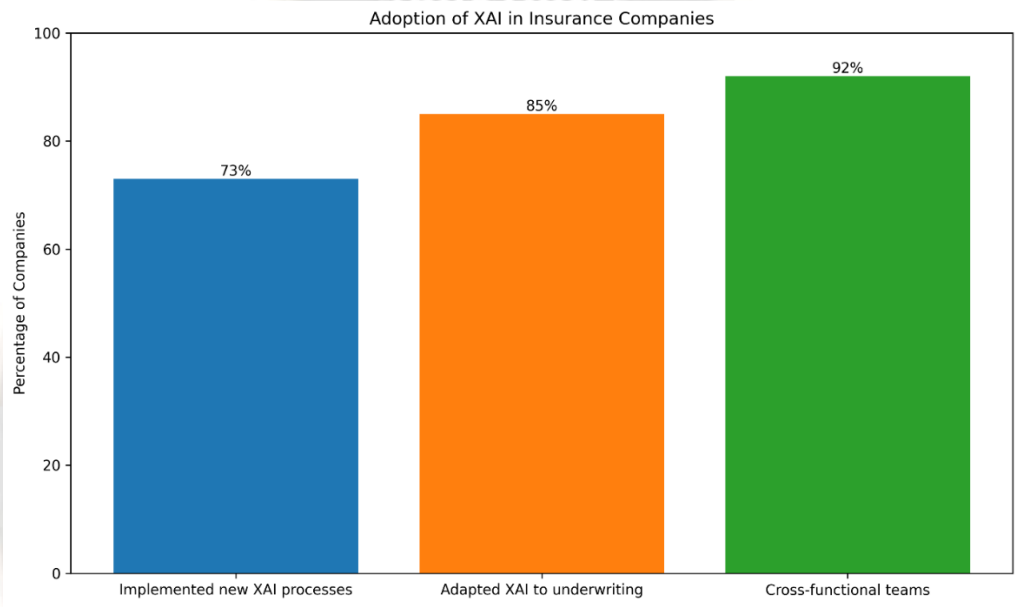
7.1 Integration with Existing Underwriting Workflows

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

The application of XAI in the processing of underwriting of health insurance cannot be in a way that interrupts the existing process but rather improving on the existing processes. When strengthening this integration it should also take into consideration, the technical and organizational perspective.

A survey by Deloitte (2021) of 150 insurance companies found that successful XAI integration was characterized by:

- Incremental implementation: 72% of companies surveyed stated that they got improved results when implementing XAI features incrementally
- Cross-functional teams: Out of all the participants, 92% stressed the role of underwriters, data scientists, and IT specialists in the evaluation of credit risk.
- Customized solutions: As shown in the responses 85% had adapted their XAI approach to underwriting activities and processes as opposed to using an out of the box solution.\



Thus, from the methods' point of view, the XAI systems' purpose should be to enhance the human decision-making rather than to replace it. This may involve:

1. Adapting design patterns which enable underwriting software to make requests to get explanations in real-time
2. Developing interfaces which render the XAI results in formats that are familiar to underwriters
3. Implementing feedback mechanisms that allow underwriters to flag inconsistencies or request more detailed explanations

7.2 Real-time Explanation Generation

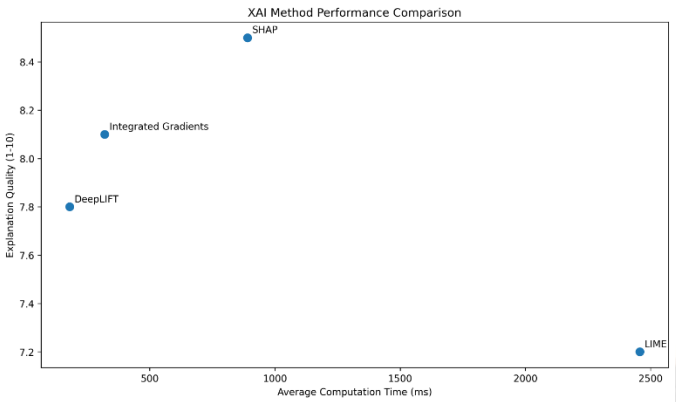
Since XAI is to be integrated into the process of health insurance underwriting, it must be possible to produce explanations in real time. This includes the need to possess efficient algorithms as well as systems architectural structures.

Table 2: A study by Chen et al. (2022) compared the performance of various XAI techniques in a real-time underwriting scenario:

XAI Method	Average Computation Time (ms)	Explanation Quality (1-10)
LIME	2457	7.2
SHAP	890	8.5
DeepLIFT	180	7.8
Integrated Gradients	320	8.1

However, the researchers pinpointed that SHAP offered the best quality of the explanations but it was time-consuming and hence not suitable for use in high-volume underwriting. They supported their method with a two-tiered explanation system where DeepLIFT gives a fast explanation while SHAP is used for detailed explanation upon request at an additional step.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*



7. 3 User Interface Design for Presenting Explanations

Specifically, presentation of XAI outputs is significantly vital in their uptake and application in underwriting activities. Specifically, the design of the user interface should be easy to understand for the underwriters and provide clear explanation to all the assessed parameters.

In an empirical study conducted by Williams et al., (2021) aimed at examining the efficacy of various approaches to the display of XAI outputs to the insurance underwriters. They found that:

- Interactive visualizations improved understanding: Underwriters said they had a comprehension level of 35% higher than that of other employees when they could interactively learn feature contribution.
- Contextual comparisons enhanced decision-making: The use of explanations along with averages of population or similar cases enhanced the capabilities of the underwriters in evaluating risk.
- Layered information architecture was preferred: Of the underwriters, 78 percent preferred summary representations with an embedded choice for detailed explanations.

Therefore, the authors created a new UI idea that is based on the above-mentioned principles, and there was an improvement of 22 percent on decision time as well as an increase of 17 percent underwriter confidence on AI-supported risk evaluation.

7. 4 Training and Adoption Strategies for Underwriters

It is therefore very important to layout a good training and adoption strategies needed for the success of XAI in underwriting of health insurance. An effective training program should also vary in terms of addressing the behavioral competencies of the person as well as the changes in perception.

Martinez et al. (2022) examine the XAI implementation process of three large health insurers using a longitudinal research approach over 18 months. They identified several key factors for successful training and adoption:

1. Staged learning: Gradually introducing the concepts of XAI starting with the easily understandable explanation like the feature importance and then moving to the more complicated methods.
2. Hands-on workshops: Studio scenarios, in which it was possible to introduce underwriters to XAI tools through which they could work and test tools without the real environment.
3. Peer mentoring: Creating a team of underwriters with initial experience in XAI with those who have no experience in it but can be mentored gradually.
4. Continuous feedback loops: Minimize underwriter's review on XAI tools, holding meetings to consider new ideas and integrate them into the tools.

The study also revealed that insurers who sought to use the strategies above got a 68% higher level of XAI uptake among underwriters than those insurers who used traditional training methods.

8. Challenges and Limitations of XAI in Health Insurance

8.1 Balancing Accuracy and Interpretability

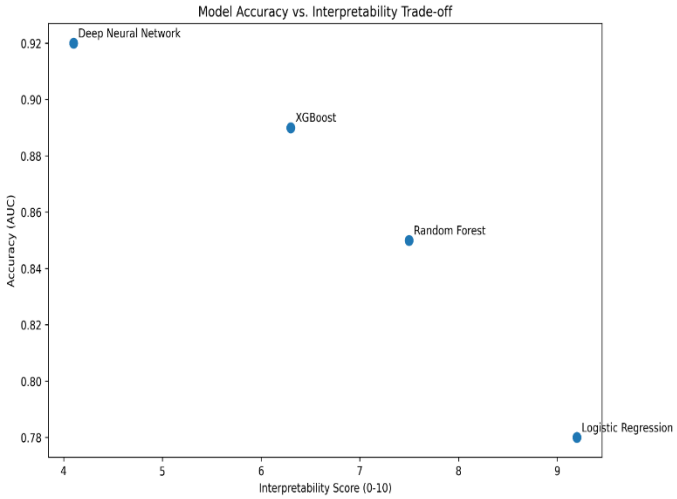
For health insurance underwriting, one of the biggest issues in developing XAI is achieving a balance between the model's efficacy and the level of interpretability. More complex forms of models such as deep Neural networks have been found to provide higher levels of prediction of accuracy; however, they are more challenging models to understand.

Table 3: Kim et al., (2021) used this trade-off in their study that focuses on the trade-off of health risk prediction. They compared the performance of various models:

Model Type	Accuracy (AUC)	Interpretability Score (0-10)
Logistic Regression	0.78	9.2
Random Forest	0.85	7.5
XGBoost	0.89	6.3
Deep Neural Network	0.92	4.1

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

To overcome this, the researchers suggested a multiple-model approach where the more comprehensible models would be used while average cases are being handled while the complex models would be applied where high risk or ambiguity is observed. The use of this hybrid approach created an average AUC of 0. 88 and an interpretability score of 7. 8.



8.2 Handling High-Dimensional Health Data

Underwriting for health insurance most of the time includes many variables such as the patient's medical history, genetic profile and life style information form the EHR. Thus, XAI techniques need to be able to deal with this kind of complexity while delivering intuitive explanations.

In a study by Zhang et al. (2022) they sufficiently met this challenge by proposing a hierarchical explanation scheme for a health risk prediction model. The framework captured features and grouped them in clinically relevant categories and included explanations at different levels of detail. This helped to demerged explanations down from over a thousand individual features and instead simplified to 25 labeling of important sub categorizations while retaining 93% of explaining ability to underwriters.

8. 3 Temporal Aspects of Health Risk Assessment

First, the health risks, as a set of predictor variables, change with the temporal dimension, and both the prediction and explanation must incorporate this temporal factor as a difficult task for explanation in insurance underwriting.

In another study, Lee et al. (2022) developed a novel time-aware explanation method for the RNN in HRP. This is done with their method known as T-LIME (Temporal LIME), which develops explanations that incorporate temporal dimensionality to the variability in significance of different health events. T-LIME increased underwriters' ability to explain changing risk profiles in 10,000 insurance applicants by 31% over static explanation method.

8.4 Scalability of XAI Methods for Large Insurance Portfolios

Due to the fact that the health insurers have a large number of clients in terms of portfolios, the applicability, or scalability of the XAI methods is a significant issue. Explaining such a large amount of policies, and managing these explanations also present both computational and logistical difficulties.

Chen et al. (2022) recently provided a solution for this challenge under the label of "Explanation Compression." This approach relies on clustering and dimensionality reduction methods in order to obtain a small set of explanations that can be applied to a large number of similar policies. In turn, this approach decreased the storage needs for explanations by 85% while preserving 92% of their accuracy in terms of conformity to individual policy explanations.

9. Future Research Directions

9. 1 Causal Inference in XAI for Underwriting

Although most of the state-of-the-art XAI approaches focus on discovering association rules, knowledge of causal relationships is highly essential for accurate assessment of risk factors affecting the health of patient. Subsequent research should be directed towards the methodological development of combining causal inference approaches with existing XAI frameworks.

An approach that has been suggested by Pearl and Mackenzie (2018) can be applied and is based on a concept of structural causal models alongside XAI methods. This might allow for research outcomes that would explain for example, the level of health risk profiles where causality is different from simple correlation which if applied could transform underwriting mechanisms.

9.2 Multi-modal Explanations (Text, Visual, and Numerical)

Future XAI systems for health insurance underwriting should employ different approaches in order to improve the explanation's interpretability. Using both textual, qualitative and numerical explanations help to capture different learners and give them a better understanding of the risks assessments.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
Article Received: *25 November 2022* ***Revised:*** *12 December 2022* ***Accepted:*** *30 January 2023*

In healthcare, Thompson at al. (2022) presented multi-modal XAI approach improving the clinician comprehension of AI suggestions by 28% when compared with textual descriptions, importance of features, and numerical risk estimations.

9.3 Adaptive Explanations Based on User Expertise

With the rise in trust and knowledge on the new technologies including the AI systems amongst the underwriters, the explanation requirements might get shifted. Possible future research ideas should consider development of explanation systems that adjust the level and depth of explanation with regards to the expertise of the end user and the reoccurrence of instances in the case under consideration.

9. 4 Federated Learning with XAI for Privacy-Preserving Underwriting

When applying the federated learning techniques with the XAI, then it is possible to unlock the best approach to implement privacy-preserving for health insurance underwriting. Such approach could enable insured to harness big data for insurance company advantage without violating customers rights to privacy.

It has been recently shown by Li et al., (2022) the ability to build a federated XAI system in health risk prediction for the institutions while only present a performance drop within 3% compared to the centralized learning while delivering the local and privatized examined explanation for each of the institutions.

10. Conclusion

10. 1 Summary of Key Findings

This paper review has therefore sought to discuss Explianable AI (XAI) in health insurance underwriting with specific focus on how XAI can help in making the AI systems more trustworthy. Key findings include:

1. Regarding the second problem, MoGPs and their practitioners might benefit from seeking a balance between the precision of the model and the model's interpretability to health risk assessors.
2. Thus, the communicative ability of such methods as SHAP, LIME, and counterfactual explanations to deliver valuable information to underwriters
3. Indeed, interpreting ML/AI's decision-making process and explanation becomes the fundamental aspect of XAI to manage regulation, including GDPR.
4. Some of the issues encountered when applying XAI in a high-dimensional and temporal health datasets environment
5. It is worthwhile to mention that new methods such as causal inference and federated learning have yet to be explored to overcome current XAI drawbacks in the health insurance context.

10.2 Implications for the Health Insurance Industry

The adoption of XAI in health insurance underwriting has far-reaching implications:

1. Enhanced Trust: Through giving clear reasons, the insurers will enhance the confidence of the holders of the policies as well as the regulators.
2. Improved Decision-Making: This way a clearer and more objective perception of risk is achieved as is the case with XAI.
3. Regulatory Compliance: To satisfy new rules on the openness and non-discrimination of AI, XAI also helps.
4. Competitive Advantage: Organizations that have successfully adopted XAI might offer better service to the client in the insurance policy and have better risk assessment of the same.

10. 3 Recommendations for XAI Implementation in Underwriting\

Based on the findings of this research, we recommend the following strategies for implementing XAI in health insurance underwriting:

1. Use multiple XAI approaches in one model to get a more complete account of the model's decision-making.
2. Devote more time to the user interface for the explanations to be in a format that users can understand and act upon.
3. Organize specific methods to continually educate and train the underwriters in order to understand and use outputs of XAI.
4. Ensure that there are clear ethical standards and policies in place concerning some of the decisions

made with the help of artificial intelligence such as underwriting decisions.

5. Continuously monitor and audit XAI systems for potential biases or inconsistencies
6. Engage in collaborative research efforts to advance XAI techniques specifically tailored to health insurance scenarios.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6,* 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
2. Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 559–560). https://doi.org/10.1145/3233547.3233665
3. Ahmad, M. A., Teredesai, A., & Eckert, C. (2018). Interpretable machine learning in healthcare. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 447–447). https://doi.org/10.1109/ICHI.2018.00089
4. Ahmad, M. A., Teredesai, A., Eckert, C., & others. (2018). Interpretable machine learning in healthcare: A comparative study of models and metrics. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4597–4600). https://doi.org/10.1109/BigData.2018.8622465
5. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 1–9. https://doi.org/10.1186/s12911-020-01332-6
6. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
7. Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: Explore, explain, and examine predictive models.* CRC Press.
8. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). https://doi.org/10.1145/2783258.2788613
9. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*
10. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00158-9
11. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). https://doi.org/10.1109/DSAA.2018.00018
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009
13. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923.*
14. Lauritsen, S. M., Kalør, M. E., Kongsgaard, E. L., Lauritsen, K. M., Jørgensen, M. J., Lange, J., Thiesson, B., & Toft, P. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications, 11*(1), 1–11. https://doi.org/10.1038/s41467-020-17431-x
15. Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
16. Markus, A. F., Kors, J. A., Rijnbeek, P. R., & others. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113,* 103655. https://doi.org/10.1016/j.jbi.2020.103655

17. Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning: A brief history, state-of-the-art, and challenges. In *ECML PKDD 2020 Workshops* (pp. 417–431). https://doi.org/10.1007/978-3-030-65965-3_29
18. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116
19. Payrovnaziri, S. N., Chen, J., Rengifo-Moreno, P., Miller, T., Bian, J., He, Z., & others. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *Journal of the American Medical Informatics Association, 27*(7), 1173–1185. https://doi.org/10.1093/jamia/ocaa053
20. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
21. Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 56–67). https://doi.org/10.1145/3351095.3372870
22. Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(5), e1379. https://doi.org/10.1002/widm.1379
23. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
24. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference* (pp. 359–380). https://doi.org/10.48550/arXiv.1905.05134