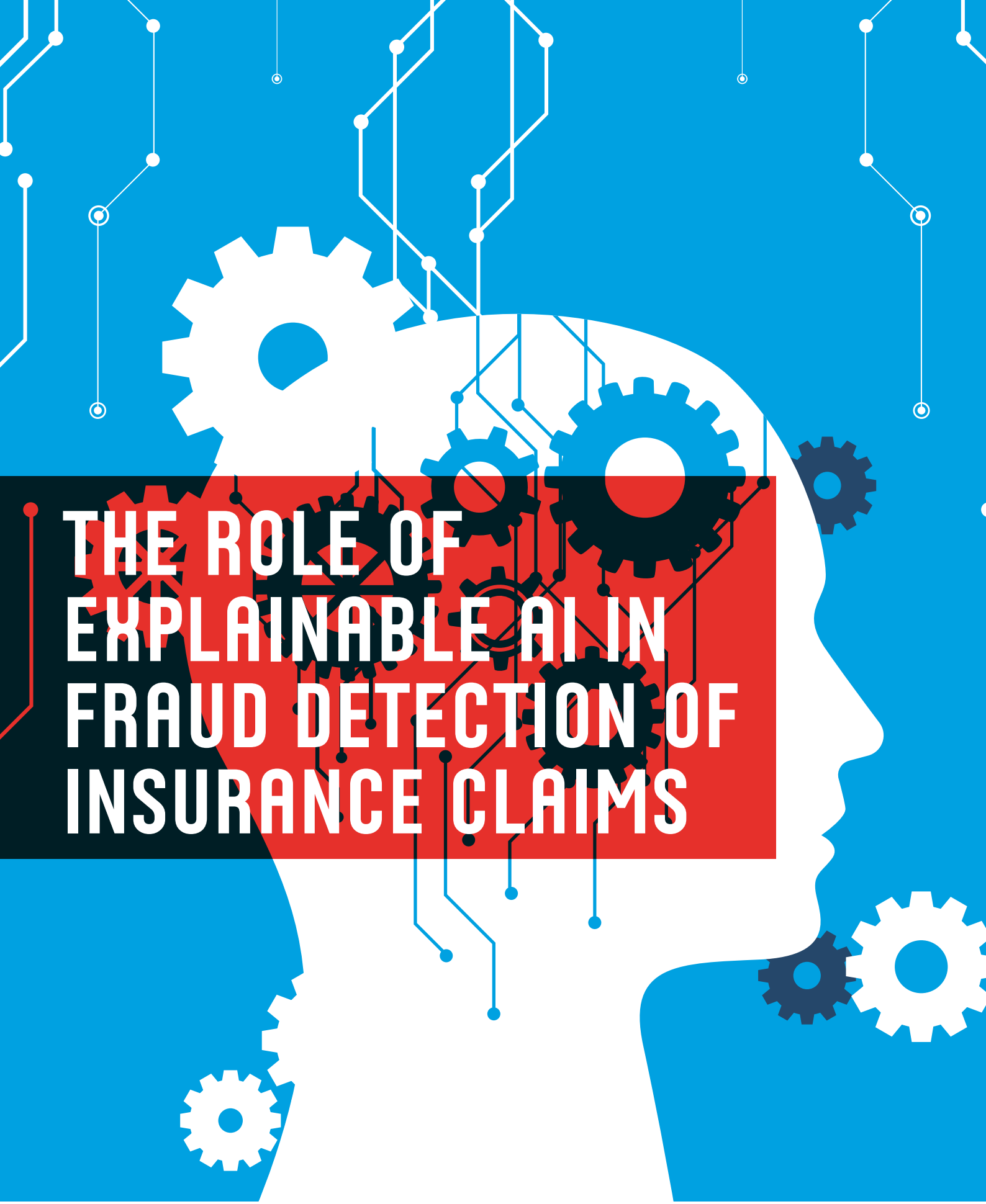

THE ROLE OF EXPLAINABLE AI IN FRAUD DETECTION OF INSURANCE CLAIMS

UNIVERSITY OF APPLIED SCIENCES UTRECHT

IMPACT YOUR FUTURE

TABLE OF CONTENTS

PREFACE

This white paper is the result of a research project by Hogeschool Utrecht, Copenhagen Business School and the Dutch Association of Insurers in the period February-June 2023.

The goal of the research project was to provide an overview of the practical implementation of artificial intelligence (AI) in fraud detection of non-life insurance claims in the Netherlands, and to investigate specifically the role of explainable AI (XAI) in this process.

The main takeaway from this research is that the implementation of AI in fraud detection is a business transformation that requires many ethical and organizational considerations. Explainability of the AI system is viewed as crucial, both from an ethical point of view (as part of the transparency principle), and from a practical point of view (as a means to gain trust and acceptance of internal stakeholders, and to form a good human-machine work collaboration). However, the practical implementation of XAI is still under active discussion and exploration in the sector.

We would like to thank the interviewees who shared their experiences with us.

Dr. Martin van den Berg
martin.m.vandenberg@hu.nl

Julie Gerlings, PhD Fellow
jge.digi@cbs.dk

Jenia Kim, MA
jenia.kim@hu.nl

Dr. Stefan Leijnen
stefan.leijnen@hu.nl

Hogeschool Utrecht - Lectoraat Artificial Intelligence
August 2023 - Version 1.0

1. INTRODUCTION

The insurance industry applies artificial intelligence (AI) in different processes [EIOPA] and acknowledges that AI must be applied in an ethical and responsible manner. Therefore, the Dutch Association of Insurers supported the industry by publishing an ethical framework [Verbond van Verzekeraars]. This framework is binding for its members. One of the requirements in this framework is that AI systems need to be transparent. The two associated standards are: 1) Before we deploy data-driven systems, we consider how we can best explain the outcomes of the system to customers, and 2) When using data-driven system, human intervention can always be called upon and an explanation can be obtained by customers about the results of an AI system. However, with certain AI systems, such as fraud detection, providing explanations is a sensitive issue and has certain challenges. According to literature, challenges include, but are not limited to, potential loss of intellectual property, gaming of the AI system, biased decisions, discrimination, and privacy and security issues [EIOPA, Coalition Against Insurance Fraud].

In other words, there is a tension between the ethical requirement for transparency and the business requirement for dealing with challenges associated with explanations. In this context, the research question we address is: How do insurers balance between the different stakeholders' need for explanations and the challenges involved in providing these explanations when using AI systems for detection of fraud in insurance claims? To answer this research question, we conducted an explorative study on the use of AI and explainable AI (XAI) in the fraud detection process of insurance claims by reviewing relevant literature and interviewing experts from different organizations.

This white paper contains the results of the research project. It generally examines the process of claim handling and fraud detection in non-life insurance claims and how AI is applied in this process. This lays the foundation for understanding the challenges of explaining outcomes of AI systems. As such the scope of this white paper is broader than to just answer the research question. Explaining outcomes of AI systems in fraud detection cannot be separated from the use of AI in fraud detection.

We define an AI system according to the OECD and EU as "a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate output such as predictions, recommendations, or decisions influencing physical or virtual environments" [OECD]. We define an AI model as a model that is developed and used in an AI system such as a decision tree, random forest, or neural network. Machine learning (ML) is the most often used type of AI system used in fraud detection. We define ML as a type of AI system that learns from data to generate output such as predictions, recommendations, or decisions. Explainable AI (XAI) is defined as: "Given a stakeholder, XAI is a set of capabilities that produces an explanation (in the form of details, reasons, or underlying causes) to make the functioning and/or results of an AI solution sufficiently clear so that it is understandable to that stakeholder and addresses the stakeholder's concerns" [Van den Berg & Kuiper].

This white paper is structured as follows.
In section 2 we present an overview of the process of fraud detection and its main stakeholders. Section 3 contains the research approach. In section 4 we present challenges of applying AI and using explanations from studying relevant literature. Section 5 contains challenges of applying AI and XAI and using explanations from practice based on the interviews with experts. In section 6 we discuss the findings of this research. Finally, section 7 contains the conclusions and a future research agenda.

2. PROCESS OF FRAUD DETECTION

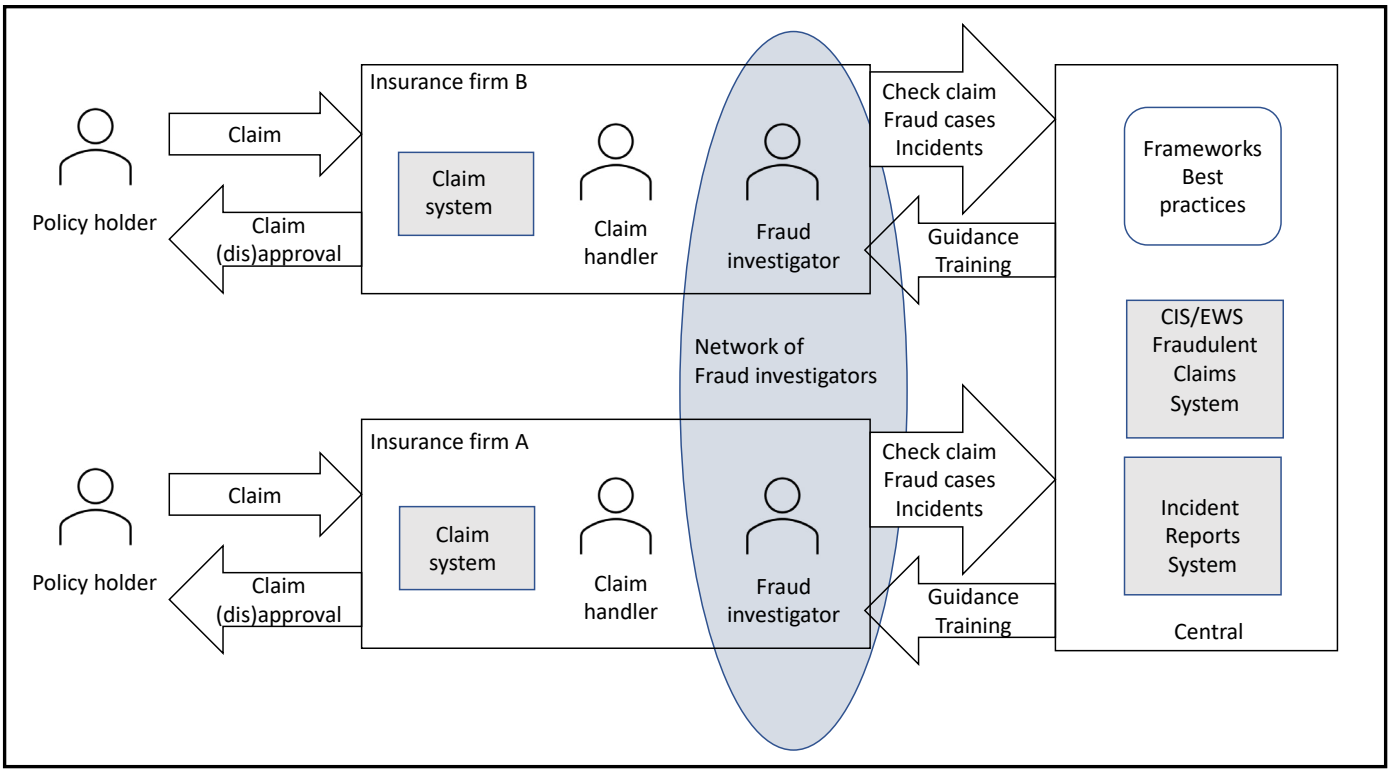

Figure 1 Process of Fraud Detection

This section contains an overview of the process of fraud detection and its main stakeholders as illustrated in figure 1. This process differs from insurance firm to insurance firm. That is why Figure 1 contains a high-level overview and does not have the pretension to be complete.

The process of fraud detection consists of the following steps:

1. A private policy holder submits a claim to the insurance firm where s/he has a policy.
2. The insurance firm processes the claim in its systems. Part of this processing is to check the claim for suspicious or anomalous information that may indicate fraud. This involves checking if the claim has been submitted elsewhere (to prevent double claiming[1]) and checking the claimant's history of insurance fraud. Some central databases, like CIS [Stichting CIS], are consulted for this purpose.
3. At this stage, the claim is either automatically approved, or manually checked by a claim handler. Some insurance firms have a partly automated system to evaluate whether the claim should go to a claim handler or to direct pay-out. Systems differ from firm to firm, even though their goal is the same; for example, some systems are fully based on business rules, while others are a combination of business rules and AI.
4. If a claim handler finds the claim of a certain level of risk or something that is out of context, s/he transfers the claim to a fraud investigator who investigates the claim in more detail. Fraud investigators operate and decide independently whether the claim is fraudulent or not. They must collect legal evidence to support or dismiss a case as fraudulent. When needed during their investigation they exchange information with the policy holder and fraud investigators of other insurance companies.
5. In the end, claims that have been found fraudulent are disapproved by the insurance company and can be reported to an External Warning System[2], which is accessible for insurers through the data platform of CIS. The insurers can report their fraud investigations and incidents to the Dutch Association of Insurers. This association also provides guidance in the form of frameworks and best practices, as well as fraud trend-analysis and alerts on modus operandi. One such framework is the ethical framework [Verbond van Verzekeraars].

1 In accordance with the Dutch civil law (article 7:960 BW): the insured person will receive no compensation under the insurance agreement if s/he would attain a clearly more advantageous position as a result.

2 This system is based on the Protocol Incident Warning System for Financial Institutions, which is approved by the Dutch Data Protection Authority.

3. RESEARCH APPROACH

The research has been conducted in a qualitative and explorative manner. The aim of the research is to gain insight into the following aspects:

1. Considerations and challenges in applying AI in detection of fraud in insurance claims.
2. Considerations and challenges in providing explanations as part of the AI application.
3. Ways in which organizations address the challenges identified in (1) and (2) in practice.

The research approach included two parts. First, a literature study was conducted to identify the known challenges that are involved in applying AI in the fraud detection process and providing explanations about the outputs of these AI systems.

Second, interviews with experts in the insurance domain were conducted to understand how the three aspects mentioned above are experienced and being dealt with in practice. We conducted five interviews with experts from four different organizations in the Netherlands, as summarized in Table 1. The interviews were conducted by two researchers and lasted about one hour each; they were recorded and transcribed. The transcripts were then coded by one of the researchers, using the NVivo software. The methodologies we build upon in our qualitative analysis are axial coding [Williams and Moser] and the Gioia method [Gioia et al.].

Table 1. List of experts

	Function	Organization	Years experience in insurance
E1	Manager Centre Against Financial Crime	A	15
E2	Chief Analytics Officer	B	19
E3	Head of Anti-Fraud	B	14
E4	Ethicist	C	7
E5	Actuary	D	18

4. INSIGHTS FROM LITERATURE

The use of AI systems in fraud detection of insurance claims is discussed in various papers. The Coalition Against Insurance Fraud reported that 56% of insurers in the US use AI as their primary mode for fraud detection [Coalition Against Insurance Fraud]. Another worldwide survey among 427 insurance professionals indicates that 42% of insurers use predictive models and AI to detect fraud [Insurance Information Institute]. Various AI technologies are being used in handling claims such as natural language processing, image analysis, machine learning, speech recognition, network analysis, web crawling, and pattern and anomaly detection [Eling et al., Coalition Against Insurance Fraud].

Several challenges are recognized in literature when applying AI systems in fraud detection:

- The first challenge is that different AI systems may be required side by side [Coalition Against Insurance Fraud, Voseler, Eling et al.]. This is partly related to the variety of data being used in fraud detection such as images, text, and tabular data [Eling].
- The second set of challenges is inherent to the insurance industry in general, and thus also to fraud detection. Insurance firms are not allowed to discriminate among policy holders. Therefore, they must prevent AI systems from algorithmic and/or human bias, and from unfairness and discrimination [EIOPA, Owens et al., Coalition Against Insurance Fraud]. When determining fraudulent claims, insurers must be very cautious not to include personal data in AI models because of the confidential and/or private nature of these data [Psychoula et al.].
- A third and typical fraud detection challenge is that fraud is not stable; it is difficult to find a "ground truth" to build AI systems on. Fraud appears in different sizes and shapes. Fraudsters change their modus operandi over time to find ways to go unnoticed [Voseler, Collaris et al., Insurance Information Institute].
- A fourth challenge is that datasets for fraud detection are imbalanced [Psychoula et al.]. The percentage of fraudulent cases is very low compared to the non-fraudulent cases. That makes it difficult to create AI systems that accurately separate fraudulent cases from non-fraudulent cases. The result is the occurrence of many false positives, i.e., claims falsely identified as fraudulent [Insurance Information Institute, Farbmacher et al.].

The above challenges impact the way these models and their outcomes can be explained:

- First, it should be noted that explainability is not a well-defined, uniform concept; rather, different stakeholders have different information requirements and therefore need different explanations [Gerlings et al., Van den Berg & Kuiper]. In fraud detection there are multiple stakeholders such as the policy holder, claim handler, fraud investigator, auditor, senior manager, and regulator. The challenge in explaining AI systems for fraud detection is to learn to know the different stakeholders and their explanation needs [Gerlings & Constantiou, Van den Berg & Kuiper].
- The second challenge is related to the situation where different AI models are used side by side. In this situation it is difficult to find the key features to create explanations since these need to be extracted from different models [Voseler].
- A third challenge is that the positive effect of providing explanations can be reversed if users perceive an explanation overload [Owens et al.].
- Fourth, and related to bias in AI systems, is the challenge that bias can also occur in explanations. In fraud detection, this is particularly relevant since decisions whether claims are fraudulent or not are made by humans. Whenever these humans get output from an AI model indicating a certain level of risk of a claim, it should carefully be considered what information to provide to these humans to minimize or avoid biased decisions [EIOPA].
- A fifth challenge is the risk of gaming the system. A detailed explanation of the features used to detect fraud might be used as a "recipe" on how to game the system, thus hindering the insurer's ability to fight and prevent future fraud [EIOPA].
- Another challenge for insurers is a combination of misconceptions on explanations. A particular explanation may be wrong (after all, the outcome of an AI system can be a false positive), a particular explanation may be confused with causality, and finally, a particular explanation may be considered as a general explanation for comparable cases [Collaris et al.].
- The last challenge has to do with available explainable AI (XAI) techniques. Many popular XAI methods, such as SHapley Additive exPlanations (SHAP), explain what features contributed to the model's output in a specific case. Such techniques provide very similar explanations for false positives and true positives, since the model classified both as positive. This is not helpful for fraud investigators, since they still waste resources on investigating false positives. Therefore, the model that classifies the claims might not be the right model to provide explanations; rather, another "meta-learning" model might be needed that can help fraud investigators distinguish false positives from true positives and allocate their resources accordingly [Zitouni et al.].

5. INSIGHTS FROM PRACTICE

This section presents the most prominent points discussed in the interviews and related to the practical use of AI and explanations in detection of fraud in claim management.

5.1 AI in Fraud Detection: The Promise

Implementation of AI systems in the claim management process is motivated by the wish to improve the efficiency and the quality of fraud detection. As evident from the excerpts below, data-driven methods such as machine learning (ML) are viewed as potentially complementary to and enhancing human expertise. The goal of this human-machine cooperation is to allow the human-experts to spend their time more efficiently by focusing on analyzing high-risk cases, instead of going through many false positives. In other words, the AI is not there to replace the human expert, but rather to enable her/him to focus on the most meaningful part of the process. The promise of ML in achieving this goal lies in the opportunity to use both more data and different types of data, compared to a manual process. In addition, ML is viewed as objective and transparent, allowing the insurers to clearly communicate their decision-making framework to different stakeholders.

"We can improve from our current way of working to a more data driven approach where we optimize the capacity that we have right now. So, the idea was not so much on how can we save on the number of people that we have in the execution at claim fraud detection, but more how can we make those people more efficient in terms of how can we detect more fraud cases that are currently not being detected by using different types of data, for instance." (E2)

"I think [ML] does bring efficiency and uniformity within a company. It does give a clear picture... Because if the computer selects for you, you have more time as a claim handler to analyze. So, you improve the quality of your process. So that's the efficiency. By giving the claim handler more time to analyze instead of selecting, will improve the quality." (E5)

"So, you can expect to have four times more efficiency in doing the manual checks in pointing towards something that in the end could be fraud. And this is what is really important because if you are a fraud investigator, you want to investigate as much of the real fraud. And, if you are constantly doing random checks, which used to be the case, the chances of you finding fraud are pretty slim." (E2)

5.2 AI in Fraud Detection: The Practice

In practice, the transition from business rules to ML-assisted systems is a transition towards the future and is a long-term investment. Such a transition demands devoted employees, good communication across departments, education of employees, interdisciplinary work, and patience. Insurance firms which do have ML incorporated in the process, use it mainly for triage (categorizing the level of suspiciousness), hence, directing suspicious claims towards further investigation and non-suspicious claims towards completion and pay-out.

To increase transparency and understanding of the ML models and how they work, models with reduced complexity are often chosen. Though most firms have been exploring the use of AI for more than 3-5 years, they chose to not implement more complex deep-learning-based models in production.

"We are using a combination of a machine learning algorithm and business rules. So, it's a combination of business rules because we know that works best." (E2)

"I do know that a student… once investigated neural networks in fraud detection but we did not apply it. That was actually too complicated. Because the model was complex, and the outcome would also lead to complexity. So, they thought it gave very nice insights, but they found applying it quite complicated. Because they didn't understand the selection of certain features." (E5)

Developing an ML model has shown to be a relatively small part of the process, whereas the human factor plays a much larger role. This includes a few aspects:

- First, the relevant stakeholders within the company need to be trained in terms of how to interact with the model and integrate its outputs in their work processes.
  "Building the model was relatively easy. To implement the model, we need to make sure that everything works at the highest standard, that we are compliant with regulations, both external and internal compliance regulations. We need to train the people." (E2)

  "Well, the most challenging part is to train the people to work with the output of the model to make them understand what the model is and how it works. You have to break these old traditions, old way of working. There are investigators who do this work for 30 years. How can

we change their minds? It's very difficult to introduce a new way of working, behavior, to look at fraud, to look at claims. That's the hardest part: to influence people, to influence their own bias. I think that's the difficult part." (E3)

- Acceptance of the model by the employees is not always easy to gain. Some people are reluctant to change their work procedures, or might not trust the model's recommendations, especially if they have not been involved in its development.

  "We also need people to accept the model. Because the human factor is that a good fraud handler could also say, well, why should I trust your model? I have 20 plus years of experience. So, we have gone a great length to implement this model." (E2)

  "…You need to involve people in an early stage into this process because otherwise they will not trust the model." (E2)

  "I would say there's always a healthy amount of people that distrust a model like this. And for me, this is not a true problem because we need people who are not in favor of the model, we want to have criticism because that keeps your model safe and keeps you sharp in taking the right precautions.... We're going to check whatever you think is not okay. And then we can see if this holds true. If it's true, we're going to fix it. If it's not true, then you learn something and you know you can trust this part of the model." (E2)

- An additional and complementary challenge is how to avoid over-reliance on the model, which might lead to biased decisions or overlooking important information and new trends.

  "The only risk of the supervised model is that new fraud trends or new fraud scenarios will not be detected by this model. So that's why the human also stays very important. And that's also why we do some training for the claim handlers. They must keep thinking themselves because there could be new fraud schemes and they have to detect them themselves... We can write new code in the model and then it can detect it, of course. But it's working together you know." (E3)

  "But they are the fraud investigators. Need to keep them fresh! We want to prevent them from trusting the model too much. We don't want to them to rely too much on the model just because the model says so." (E2)

- Finally, it is crucial to establish an ongoing, flexible, and direct feedback loop between the developers and the end-users of the model. Such cooperation enables a continuous two-way learning process. On the one hand, the model can be improved based on the experience and contextual knowledge of the users (claim handlers and fraud investigators). On the other hand, using the model can help the investigators to keep an open mind, avoid tunnel vision, and consider new types of information that they might not have noticed otherwise.

  "It could be still that a claim handler detects fraud in a claim and the model didn't show a high or medium score. We need this feedback. Why did the model miss it? Maybe some low scores should be higher? Or is it a new fraud scheme and do we need to add information to our model?" (E3)

  "Sometimes fraud investigators keep on digging, try to find something that could indicate a possible fraud. That is not the way we want to work. You have to prevent tunnel vision. It is very important that we change the working method of our fraud investigators and especially the old-fashioned fraud investigators who are doing their job for many years on the same traditional way. This requires a change of behavior. We must change their vision on how to find fraud (supported by data, AI and ethical framework). And yeah, that's the way to maybe prevent or influence their bias." (E3)

5.3 Ethical Considerations

An articulated and deep understanding of the ethical challenges in the process of fraud detection in general, and ML-assisted fraud detection in particular, was seen across all firms. A considerable effort in applying the ethical framework [Verbond van Verzekeraars] has been made by all companies interviewed. The focus on implementing the ethical framework is immense, the leading principle being a human-centered and transparent system, even at the cost of missing some fraud cases.

"[Explainability is] more important than developing a tool to detect everything. You can better miss some fraud than make a mistake on that last one. What I mean is you can better be ethical, secure, explainable, and transparent, and therefore miss some fraud, than to use data or methods to detect more fraud, but be less ethical, secure, explainable, and transparent. That's my opinion…" (E3)

5.3.1. Ethical Framework

The leading guideline used by the companies is the ethical framework of the Dutch Association of Insurers [Verbond van Verzekeraars], which is binding for the association's members. This guideline is inspired by national and EU laws and regulations, with a more rigorous approach at times.

"In the ethical framework it says even if the National law or the European law allows something, and the ethical framework of the Insurance Association says no, we do not do that.

We ask our members to follow the rules of the ethical framework, so we narrow our own boundaries, even if there's more possibilities within the (European) law." (E1)

"We have a strong ethical framework, which is a nine-page legal document which explains to what type of things a model should adhere to. These consist of the seven principles of trustworthy AI from the high-level expert group of the EU that published this paper." (E2)

The ethical framework is based on the "Ethics guidelines for trustworthy AI" [European Commission] which contains seven principles that AI systems should meet in order to be deemed trustworthy:

- Human agency and oversight.
- Technical robustness and safety.
- Privacy and data governance.
- Transparency.
- Diversity, non-discrimination, and fairness.
- Societal and environmental wellbeing.
- Accountability.

These principles were mentioned multiple times during the interviews, with special focus on accountability, safety, transparency, non-discrimination, and human agency.

Accountability and safety

The interviewed companies prefer developing their AI solutions in-house, in order to have full control and full accountability. They stress the importance of ensuring that the model is robust and safe and continually testing to see if it needs to be updated or retrained.

"… and the main reason for that [developing in-house] was to be in control yourself. To ensure we comply with our ethical framework, law and legislation." (E3)

"Everything is being tested and tests are being tested and over and over again… sometimes we retrain the model based on the outcome of tests and also, we did some shadow runs. We run the model for quite some time to do it in parallel but not in production and see what would be the performance." (E2)

Transparency

Transparency towards internal stakeholders is regarded as very important, mainly as a means to gain employees' trust and acceptance and improve their understanding of the AI system; this is discussed in detail in Section 5.4.

There is, however, a sensitive aspect to the transparency principle, which has to do with how much information can and should be disclosed to the customer. On the one hand, the companies have a moral (and sometimes legal) obligation to disclose the use of an algorithm in their fraud detection process and to explain what the algorithm does. On the other hand, full transparency about proprietary in-house algorithms is problematic in terms of competition between firms, and it also creates a risk of gaming the system.

"Yeah, for me, it's just a question like you said, what stakeholder demands, what type of explanation. And that's just an uncertainty. So we'll have to wait and see what type of story people want to hear about why they're affected in a certain way. I do sense that a lot of internal more business minded people are hesitant to just give away the formula. So if, say, a client says, I want to know exactly how my decision was made and you have sort of proprietary AI tech or a model that determined this outcome, are you going to be forced to give away the proprietary formula that led to the outcome? Or can you sort of redact it in a certain way to only give the relevant factors? So I sense more hesitation internally from people that build the models that maybe put a lot of time and money and effort into building them. So that would be more like a risk in terms of competition, but for me, not a risk in terms of explainability. I would say whatever level of explanation a client wants, give it to them. If they're a mathematician and they'd love to see how the model made this calculation. Sure. Why not?" (E4)

Moreover, disclosing full information to the customer during an ongoing fraud investigation is especially risky, since it might affect the investigation and its outcomes.

"… at the moment you start a fraud investigation, you don't want to tell them all you are doing and what you do in fraud detection and fraud investigations. That can be important to preserve any fact finding you want to be doing. For instance, the truth. They'll be getting in some kind of a danger zone when the subject knows they are being investigated. So, there is the problem with explainability and transparency versus the research on fraud cases." (E1)

"They must inform clients when they are processing their personal data. But that will not mean you have to tell them all the details of what you're doing in your process. But you must explain why something is taking up a little bit more time before they're getting a decision on their claim, for instance. But it's always difficult." (E1)

Non-discrimination

There is a general trend in the interviewed companies to prioritize the clients and their experience rather than solely focusing on detection of more fraud. Using ML to identify suspicious claims and ending up wrongly accusing someone of fraud can have tremendous consequences to the individual. Moreover, it can tear the image of a company and the entire industry down. Therefore, firms have high standards for what data is being fed into the models to minimize the risk of discrimination or bias towards specific groups. For example:

"For the detection of fraud, area codes are a no go. You cannot create any fallout of your straight through process just on an area code. I do know that you can use it for risk management, for risk evaluations. And for instance, my car insurance premium is a bit lower than two zip codes to my left. But that's a risk assessment issue and not a fraud assessment issue." (E1)

Human agency

Human oversight is another crucial aspect of implementing ML in the fraud detection process. The model is used only to assess the risk and output a score; the rest of the process, which includes the investigation and the final decision, is always performed by a human expert. This is also expressed in how the model is being named and talked about, and it is part of ensuring the intended use of the model. As the quote below shows, there is a deliberate distinction between 'fraud detection' and 'fraud risk', which emphasizes that it is the investigator, and not the model, who detects fraud.

"We call it a fraud risk model because the model itself doesn't detect fraud. It's always the human who must assess this risk and must decide if it is a possible fraud or not. An important thing in the development of our tool was a human in the loop. So first, the system presents to the claim handler, these are the fraud risks identified. Then the claim handler must look at it and must assess these risks. He might ask some questions to the client or ask for additional information …and then he says, well, I don't trust this claim to be valid. Maybe it's fraud. Then it goes to the fraud investigator. And then he also looks at it. Are there enough indicators for fraud? If so, okay, we take over this claim and start a fraud investigation. The investigation has to point out if it is possible fraud or not. So, it's a human who always makes the decision." (E3)

5.3.2 Implementation of Ethical Framework

Ethical guidelines have been incorporated in different ways at the firms interviewed. What they have in common is that the incorporation of the framework is thorough and well thought about. The interviewed companies have typically started out with workshops to create awareness about the framework and guidelines in general.

"I've done some ethical workshops with our fraud department. And as we were implementing the ethical framework internally, we've looked at the fraud detection process within [Company] to see if there's any risks involved that touch upon points from the ethical framework." (E4)

However, awareness is not sufficient when it comes to building responsible AI. The data scientists who actually work on developing and iteratively testing the model need practical instructions that translate the ethical principles into actionable tasks.

"…but if you're a data scientist, you want to have something much more practical. So, we created an AI assessment that covers all the seven principles in the ethical framework, but in a questionnaire type of way. It asks you what type of data are you going to use? Does it contain [personal identifiable information]? And if so, is your data protection officer involved? And did he or she check the "baseline for data processing?" (E2)

Moreover, it is not a one-time assessment; every iteration of the model demands a review of the data used and a possible update of the checklist.

"…the assessment starts and ends basically never because once it is in production, you also need to come back to the assessment every six months or every year, depending on the type of use case that you're doing, and you need to update this document." (E2)

One example of the complexity of practically implementing ethical guidelines is how to eliminate discriminatory features from the data going into the model. The basics of supervised ML start with learning from historic data and build upon that to establish a probability of a new claim falling into one of the categories. The features going into the AI-model are carefully chosen to minimize the risks of discrimination and biases. Therefore, some features, such as 'country of origin' or 'nationality' might be excluded or altered before they go into the model. However, some features are less easily identified as problematic, as they do not seem discriminatory by themselves, but they do serve as a proxy for a discriminatory feature.

"We're putting a lot of effort in bias detection. We created some tools ourselves to detect whether there is a statistical bias for the model to affect certain people which are vulnerable. So, either based on a religion, sexual orientation, a social class… there are 25 attributes that are prohibited to use because they are discriminatory. These are clear for everyone. The true harm is in the proxies of those 25 attributes. So, we are now in a late phase of deploying also this bias detector based on features that might be a proxy to discriminatory features." (E2)

In the example presented by the interviewee, it turned out that even though 'country of origin' was excluded from the data, there was a proxy feature for this information hidden in the 'marital status' feature, since one of its values was 'Married outside of the Netherlands' (a proxy for a foreign country of origin). This was discovered by a dedicated bias detection tool built by the company.

"If you are married, you both take a mortgage. It has some impact on the product. So, you are allowed to ask that: married? Yes or No. But in this case the bias detector discovered that there was a strong proxy in this marital status attribute. And it was just because we had different categories in this attribute. It could be Yes, it could be No, but it could also be Yes, married outside of the Netherlands, which was a different category, which was not being used by us deliberately in the model. But marital status was part of the model. And now, potentially this could be a proxy for ethnical background... So, we did a recode of this attribute to simple Yes or No." (E2)

5.4 The Role of Explainable AI

As mentioned above, transparency about the AI system towards internal stakeholders is regarded as an important ethical principle. One such internal stakeholder is the managerial level, for whom it is important to understand how the model works, as they need to sign off on it and therefore are accountable for it. This need for transparency and explainability often drives the preference towards less complex, but more explainable, models.

"When I look at the senior managers and directors, they also want to understand. They tend very much towards the less complex models for the time being. Maybe in time it will change. Yes, but for the time being when I look at it and I see how the people at the top think, I think they are quite careful…" (E5)

Another important stakeholder is the internal end-users of the model, i.e., the claim handlers and the fraud investigators. Since they need to work with the outputs of the model, they need to understand what these outputs are. In addition, they need to be prepared to answer questions about these outputs from the customer, if such questions arise.

"Before we started this, we assessed all risks. And there's one risk we described. We must be clear what the outcome of this model means. We must be clear that the claim handler must understand, but also the fraud investigator must understand how this model works and what they are seeing." (E3)

"… you can explain the model well, but it is sometimes too in-depth for the claim handlers. That's why I come up with competence, to be able to understand such a model properly. For the current colleagues who work in claims, they have learned things in a different way… But because they do not yet have that competence, they need to get to know those AI models well, but they also need to know how the score is arrived at… if the claim handler does not understand why he is asking for certain information and the customer asks, why are you asking this? Yes, then it will be difficult. So, you actually need some kind of further training... The customer wants a good explanation." (E5)

However, there is a tension between the pros and cons of providing detailed explanations about the outputs of the model. Some companies provide the claim handler with the risk score outputted by the model, as well as a detailed explanation in natural language about the features that contributed to this score. The advantage of this approach is that it gives the claim handler an indication on what is suspicious in the claim and where s/he should look first.

"Very important thing we built in. So, the model, of course, gives a score. But to the person who receives the claim, there's an explanation. You received this claim to be handled manually because XYZ and then it gives the explanation in human language…For instance, a highly unusual price for a claim like this or a combination of certain factors. This same claim amount has been issued before, or an email address or this bank account was used in a similar claim before, but with another policyholder... So, there are different rules in the claim process." (E2)

However, this level of explainability also has some potential disadvantages, as it might create a bias or a tunnel vision of the handler. Therefore, some of the interviewed companies chose not to provide detailed explanations; instead, they order the cases by levels of risk, so that the most suspicious cases are handled first, but they expect the handlers and investigators to do the investigation "from scratch" to avoid potential bias by the model.

"So, it might give a score to a certain case and that case might be prioritized. And then the human comes in and starts to do their own research... we talked about in our explainable AI workgroup, how important it is for the human not to just see all the factors that the AI has determined as fraudulent because that might already bias them in a certain direction. It might already color their judgement." (E4)

"So, what happens now is they'll get a file, but it'll be a clean slate. It'll be open research. So, I mean, obviously they'll be biased in knowing this file has been registered as possible fraud, but it won't give a very determinate score, or all the factors involved. So, the human that does the fraud research sort of just starts from scratch…" (E4)

The level of explainability in models such as random forest or boosting models (XGBoost) may seem simple on a general level, however reasoning through the decision from a single claim evaluation can be very difficult. Therefore, firms have introduced SHAP and LIME as explainable components in their model framework. These explainable frameworks can assimilate an instance (case) and show which features are most likely to have the highest impact on the evaluation.

"We use a relatively easy simple machine learning algorithm where you can get quite good results with SHAP or LIME with it." (E2)

In combination with simpler models that do not involve deep learning, firms overcome the challenge of extracting information about the reasoning of the ML models choices. Now, the challenge is to ensure understanding from the stakeholders who need the information.

"I talked with a colleague who also worked with these models, and he said yes, you can explain the model well, but it is sometimes too in-depth for the claim handler…" (E5)

Though SHAP and LIME plots have been extensively promoted as explainable and interpretable, they still cause confusion to many stakeholders outside the data science domain since they are not contextual to the people who receive them. Moreover, claim handlers and fraud investigators tend to be analytical people who seek information until they understand in detail what is going on. Therefore, the plots can be too detailed, or may show the wrong context, to be useful for these

stakeholders. One firm has generated indicators based on the plots, which are formulated in natural language to overcome this challenge.

"So, the claim handler sees on his screen the claim. Based on our model the claim gets a risk score of High, Medium or Low. Our model will also add a simple explanation in three to five lines. So not just red, orange, green or a difficult explanation or code, but explanations like: 'watch this invoice or look at this address, it's known in another case. See claim number x'. So, the data scientists must make a translation from the code to send it to the claim handler to make it clear for them how to interpret this risk." (E3)

Furthermore, the concern of biased judgement appears again in the explanations, if not addressed in the evaluation of data and quality assurance.

"… if there's bias in the data, there will also be bias in the model. So potentially, yes, if you do not check against bias, there could potentially be also bias in the explanation." (E2)

The difference here is that there are different stakeholders interacting with the explanation than with the data. Data scientists and engineers have the knowledge and expertise required to detect and eliminate bias, whereas, if bias is transferred to the explanation, it is presented to people without a background in data science or related fields. The stakeholders in claim investigation and fraud detection are experts in their own fields, which can make it difficult to ensure they use the information provided by the model in the best way. Therefore, most of the firms have ensured different ways of cross disciplinary knowledge sharing and feedback loops to ensure that both the model and the humans are in the best state to collaborate.

6. DISCUSSION

Our research provides an up-to-date overview of the practical use of AI in fraud detection of insurance claims in The Netherlands. Based on the five interviews we conducted, we conclude that:

- Insurance firms recognize the potential benefits of integrating AI systems into their fraud detection process.
- They acknowledge the limitations of the technology and determine its place in the whole process accordingly, so that the cooperation between the model and the human experts is optimal.
- The implementation of AI is taken seriously: it is a long process, and a lot of effort is put not only in the technical aspects but also in the human and organizational aspects.
- There is a lot of awareness of the ethical principles that need to be met to implement AI responsibly. The Dutch Association of Insurers provides a clear ethical framework, which is based on EU guidelines. The ethical framework is binding for the association's members. Translation of the ethical framework into operational and actionable instructions is done in-house by each company.
- Explainability of the AI system is viewed as crucial, both from an ethical point of view (as part of the transparency principle), and from a practical point of view (as a means to gain trust and acceptance of internal stakeholders, and to form a good human-machine work process). However, the practical implementation of XAI is still under active discussion and exploration. Specifically, it is still not clear (a) how much do we want to explain to each stakeholder (management, claim handlers, fraud investigators, customers), (b) what would be a meaningful explanation for each stakeholder (level of detail, format of presentation, etc.), (c) how do explanations affect the work process and the decision-making process (create bias and tunnel vision, help to focus and prioritize, etc.). It seems that at this point in time, i.e., after 3-5 of years of exploring integration of AI-based systems in their fraud detection processes, the companies have reached the phase when these questions are ready to be addressed.

Comparing the insights from practice with the insights from literature regarding XAI, we see that the challenges identified in literature are partly different from the ones mentioned in the interviews. What both have in common is the risk that models and thus explanations can be biased and can discriminate. In fraud detection, humans ultimately decide and when these humans get the output from the model, it should carefully be considered what information to provide to these humans to minimize or avoid biased decisions [EIOPA].

The difference between literature and practice is that there are risks mentioned in the literature that do not (yet) occur in practice, such as the use of different models side by side [Voseler], the explanation overload [Owens et al.], the risk of gaming the system [EIOPA], misconceptions on explanations [Collaris et al.], and the risk that the model that is used to flag a claim is not the right model to provide explanations [Zitouni et al.]. This can be attributed to the fact that application of AI in fraud detection is relatively new in the Netherlands, and the desired level of explainability is still being under active discussion and exploration; therefore, practical implementations of XAI are still few and in an early stage. It can be expected that once experimentation with XAI solutions is in a more advanced stage, the challenges described in the literature will arise as well.

This study has limitations. First, the results are based on only five interviews with representatives of organizations in the Netherlands. Second, interviewees may be biased on their perception of the firm's practices. And lastly, the interviewees all belong to the managerial levels in the organizations, and might not fully represent the challenges encountered by the employees who interact with the AI systems in practice (such as developers and end-users).

LESSONS LEARNED

Implementing ML models
Developing ML models for fraud detection is not the hard part, but implementing the model is.

- Use a multidisciplinary approach and cooperation between teams.
- Invest time in education and training.
- Involve stakeholders in early stages.

Ethical considerations
There is a lot of awareness among the interviewees that AI needs to be applied in a responsible way.

- The ethical framework [Verbond van Verzekeraars] is a good starting point for insurers to develop their own practical ethical guidelines.
- Accountability, safety, transparency, non-discrimination and human agency are top priorities in the process of AI implementation.
- Transparency in general, and explainability in particular, can be tricky to implement, when taking into account all pros and cons.

7. CONCLUSIONS

The main takeaway from this research is that the implementation of AI in fraud detection is a business transformation that requires many ethical and organizational considerations. Education and inclusion are crucial to ensure a successful integration of AI into the fraud detection process, and an optimal human-machine cooperation.

Interviewees are well aware of the risks, limitations and challenges of applying AI and insurance firms have ethical frameworks in place to mitigate these risks. Explainability of the AI system is viewed as crucial, both from an ethical point of view (as part of the transparency principle), and from a practical point of view (as a means to gain trust and acceptance of internal stakeholders, and to form a good human-machine work process). However, the practical implementation of explainable AI is still under active discussion and exploration in the sector.

The implementation of AI in fraud detection is certainly an area for future research, especially the way humans and machines cooperate, e.g., what is the optimal human-machine collaboration to reduce bias in judgements?

8. REFERENCES

- Verbond van Verzekeraars. 2020. Ethisch kader. Retrieved from https://www.verzekeraars.nl/branche/zelfregulering-soverzicht-digiwijzer/ethisch-kader-datatoepassingen
- EIOPA (European Insurance and Occupational Pensions Authority). 2021. AI Governance Principles towards Ethical and Trustworthy AI in the European Insurance Sector. Retrieved from https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa-ai-governance-principles-june-2021.pdf
- OECD. 2022. Recommendation of the Council on Artificial Intelligence. Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
- Van den Berg, M., & Kuiper, O. 2020. XAI in the financial sector: a conceptual framework for explainable AI (XAI). Retrieved from https://www.internationalhu.com/research/projects/explainable-ai-in-the-financial-sector
- Coalition Against Insurance Fraud. 2020. Artificial Intelligence and Insurance Fraud. Retrieved from https://insurancefraud.org/wp-content/uploads/Artificial-Intelligence-and-Insurance-Fraud-2020.pdf
- Stichting CIS. Retrieved from https://stichtingcis.nl/en-us/
- Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. International Management Review, 15(1), 45-55.
- Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. Organizational research methods, 16(1), 15-31.
- Insurance Information Institute. 2022. Insurance fraud report 2022. Retrieved from https://www.friss.com/downloads/insurance-fraud-report-2022
- Eling, M., Nuessle, D., & Staubli, J. (2021). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. The Geneva Papers on Risk and Insurance-Issues and Practice, 1-37.
- Vosseler, A. 2022. Unsupervised Insurance Fraud Prediction Based on Anomaly Detector Ensembles. Risks, 10(7), 132.
- Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J., & Castignani, G. 2022. Explainable Artificial Intelligence (XAI) in Insurance. Risks, 10(12), 230.
- Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. 2021. Explainable machine learning for fraud detection. Computer, 54(10), 49-59.
- Collaris, D. A. C., Vink, L. M., & van Wijk, J. J. 2018. Instance-level explanations for fraud detection. 5th Data Science Summit (DSSE 2018), (5).
- Farbmacher, H., Löw, L., & Spindler, M. 2022. An explainable attention network for fraud detection in claims management. Journal of Econometrics, 228(2), 244-258.
- Gerlings, J., Shollo, A., & Constantiou, I. 2021. Reviewing the Need for Explainable Artificial Intelligence (xAI). In 54th Annual Hawaii International Conference on System Sciences, HICSS 2021 (pp. 1284-1293). Hawaii International Conference on System Sciences (HICSS).
- Zitouni I., Postema, J.T., Sznajder, D. & Van Es, R. 2022. Explainable AI in Fraud Detection. Retrieved from https://www.milliman.com/en/insight/explainable-ai-in-fraud-detection
- European Commission. 2016. Ethics guidelines for trustworthy AI. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Gerlings, J., & Constantiou, I. 2023. Machine Learning in Transaction Monitoring: The Prospect of xAI. In T. X. Bui (Ed.), Proceedings of the 56th Hawaii International Conference on System Sciences (pp. 3474-3483). Hawaii International Conference on System Sciences (HICSS).