Towards True Explainable Artificial Intelligence for Real World Applications

Hani Hagras[a]
The Computational Intelligence Centre, School of Computer Science and Electronics Engineering, University of Essex, Wivenhoe Park, Colchester, CO43SQ, U.K.

Keywords: Explainable Artificial Intelligence, Fuzzy Logic Systems.

Abstract: We are entering a new era which is characterized by huge amounts of data which are generated from almost every application in our everyday lives. It is getting easier to organise such huge amounts of data via efficient data bases and ever growing and cheaper data storage systems (which can nicely scaleup in cloud based solutions). Due to the huge sizes, high dimensionality and complex relationships of such data, Artificial Intelligence (AI) technologies are well placed to handle such data and generate new services, business opportunities and even provide breakthroughs to completely change our lives and realise new industrial revolution as anticipated. The vast majority of AI technologies employ what is called opaque box models (such as Deep learning, Random forests, support vector machines, etc) which produce very good accuracies but it is quite difficult to analyse, understand and augment such models with human experience/knowledge. Furthermore, it is equally difficult to understand, analyse and justify the outputs of such opaque AI models. Hence, there is a need for Explainable AI (XAI) models which could be easily understood, analysed and augmented by the users/stake holders. There is a need also for such XAI models outputs to be easily understood and analysed by the lay user. In this paper, we will review the current trends in XAI and argue the real-world need for true XAI which provides full transparency and clarity at the model and output level.

1 INTRODUCTION

Over the past few decades, Artificial Intelligence (AI) has moved from the realms of science fiction to become a key part of our day-to-day lives and business operations. A report from Microsoft and Ernst and Young (EY) that analysed the outlook for AI in 2019 and beyond, stated that "65% of organisations in Europe expect AI to have a high or a very high impact on the core business." (Chavatte, 2018).

In the banking and financial industries alone, the potential that AI has to improve the sector is vast. Important decisions are already made by AI on credit risk, wealth management, financial crime, intelligent pricing, product recommendation, investment services, debt-collection, etc.

The adoption of AI across business sectors has not come without its challenges. In a recent forecast (Press, 2019), Forrester predicted a rising demand for transparent and easily understandable AI models, stating that "45% of AI decision makers say trusting the AI system is either challenging or very challenging.". This isn't very surprising when we consider that most organisations today still work with what are known as "opaque box" or "black box" AI systems. These opaque models rely on data and learn from each interaction, thus can easily and rapidly accelerate poor decision making if fed corrupt or biased data. Such "black box" AI systems also leave the end customer in the dark, doing nothing to instil trust in the technology. This lack of trust is also being compounded by widespread scepticism from consumers who are reticent to share their personal data, especially if they cannot be sure how it is going to be used.

Fortunately, Explainable AI (XAI) models have the capabilities to overcome the abovementioned concerns, while providing reassurance that decisions will be made in an appropriate and non-biased way.

In this paper, we will present various XAI approaches while arguing the case for the need for True XAI systems for real world applications which are characterized by models which could be easily

[a] https://orcid.org/0000-0002-2818-5292

Hagras, H.
Towards True Explainable Artificial Intelligence for Real World Applications.
DOI: 10.5220/0012272500003595
In *Proceedings of the 15th International Joint Conference on Computational Intelligence (IJCCI 2023)*, pages 5-13
ISBN: 978-989-758-674-3; ISSN: 2184-3236

analysed, understood and augmented by the relevant stake holders. Also the outputs of these models should be easily understood and analysed by the lay user.

Section 2 provides an overview on XAI approaches. Section 3 provides a discussion by what we mean by "True" XAI. Section 4 provides a review of some real world deployments of such True XAI. Section 5 provides the conclusions and future work.

2 OVERVIEW OF XAI APPROCAHES

XAI systems are expected to be highly transparent models which explain, in human language, how an AI decision has been made. Ideally, they do not solely rely on data, but can be elevated and augmented by human intelligence. These systems are supposed to be built around causality, creating space for human sensibility to detect and ensure that the machine learning is ethical and course-correct if it is not. This is extremely valuable when we consider that most industries don't usually have the privilege of finding out that their AI model is biased until it's too late.

In many sectors of the economy, XAI is creating positive outcomes for both the industry and the customer. For example, in banking and finance, XAI systems have allowed institutions to carve out new revenue streams. By providing insights into a particular AI outcome, banks can reroute customers that have been denied a service and recommend a more suitable option for them for which they would qualify. This allows banks to provide highly personalised services to customers and explore new product lines based on evidenced demand. The customer, on the other hand, receives an explanation of why a particular service has been denied and an alternative is offered in its place. With this insight, the customer may also be able to make lifestyle changes in order to attain their financial goals and improve in their financial wellbeing.

Figure 1 depicts a summary as provided by (Gunning, 2017) showing some AI techniques performance vs explainability where it is shown that black box models like Deep Learning give best prediction accuracy vs Decision Trees which provide higher explainability contrasted by prediction accuracy.

XAI can also be categorized according to different criteria:

- **Intrinsic or Post Hoc:** whether the model itself is architecturally explainable (transparent model), or the technique tries to explain an opaque model.
- **Result of the Interpretation:** how is the "interpretation" returned to the user?
    - Feature summary statistic.
    - Model internals (weights).
    - Data point analysis related to the model.
    - Surrogate model.
    - A set of linguistic and numerical explanations.
- **Local or Global Explanations.**
- **"Reference Based" and "Non-Reference Based":** whether you need an example/reference to provide an explanation.
- **Level of Interpretability:** do I need technical knowledge or not? If so, how much?

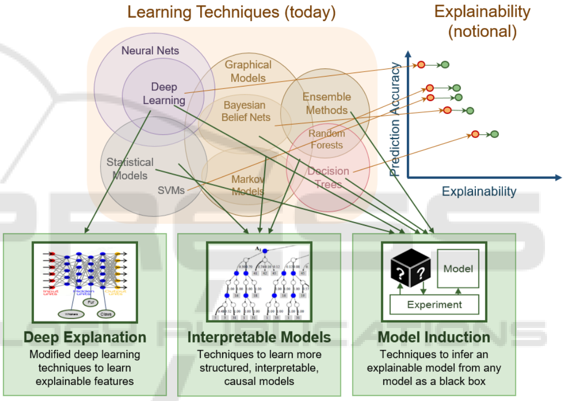

Learning Techniques (today)
Explainability (notional)
Neural Nets
Deep Learning
Graphical Models
Bayesian Belief Nets
Ensemble Methods
Random Forests
Statistical Models
SVMs
Markov Models
Decision Trees
Prediction Accuracy
Explainability
Model
Experiment
Deep Explanation
Modified deep learning techniques to learn explainable features
Interpretable Models
Techniques to learn more structured, interpretable, causal models
Model Induction
Techniques to infer an explainable model from any model as a black box

Figure 1: Existing AI techniques- Performance vs Explainability (Gunning, 2017).

As shown in Figure 1, in (Gunning, 2017), they suggest various approaches to realise XAI, the first approach applies to Deep Learning and Neural Networks (which according to Figure 1 and (Gunning, 2017) have the highest predictive power) which is termed as deep explanation. This approach tries to process the deep learning (or neural network) techniques to learn explainable structures. Some examples of such techniques can be found in (Montavon et al., 2018) including, the Layer-wise Relevance Propagation (LRP) technique (Bach, 2015).

The second approach to XAI in Figure 1 is called interpretable models which are techniques to learn more structured and interpretable casual models which could apply to statistical models (e.g. logistic regression models, naïve bayes models, etc), graphical models (such as Hidden Markov Models, etc) . However, like the deep explanation techniques, the output of these models could be analysed only by an expert in these techniques and not by a lay user.

The third XAI approach is what is termed model induction which could be applied to infer an interpretable model from any black box model (Gunning, 2017). According to (Ribeiro, 2016a), although it is often impossible for an explanation to be completely faithful unless it is the complete description of the model itself, for an explanation to be meaningful it must at least be locally faithful, i.e. it must correspond to how the model behaves in the vicinity of the instance being predicted. As mentioned in (Ribeiro, 2016a), local fidelity does not imply global fidelity: features that are globally important may not be important in the local context, and vice versa. While there are models that are inherently interpretable, an explainer (or model induction) should be able to explain any model, and thus be model-agnostic. An interpretable explanation need to use a representation that is understandable to humans, regardless of the actual features used by the model. In (Ribeiro, 2016a) a method was presented to explain a prediction where they used sparse linear explanations, which lack the explanation of the interconnection between the various variables driving the given decision.

In (Ribeiro, 2016b), they introduced Anchor Local Interpretable Model-Agnostic Explanations (aLIME) which is a system that explains individual predictions with crisp logic IF-Then rules in a model-agnostic manner. Such IF-Then rules are intuitive to humans, and usually require low effort to comprehend and apply (Ribeiro, 2016b). However the IF-Then anchor model presented in (Ribeiro, 2016b), use crisp logic and thus will struggle with variables which do not have clear crisp boundaries, like income, age, etc. Also the approach in (Ribeiro, 2016b), will not be able to handle models generated from big number of inputs. Also, another major problem in an anchor approach, is the inability to understand the model behaviour in the neighbourhood of this instance and how the prediction can be changed if certain features could be changed, etc.

Another very important XAI model induction is based on Shapley values (Sundararajan and Najmi, 2019) which are used within various AI platform. However, Attributions depend on baselines: “baseline” is the word they use for a “reference instance”. The values of each attribution and interpretation thereof depend entirely on the choice of baseline, it's as important as knowing what questions to ask when seeking an explanation. One must never omit the baseline from any discussion of the attributions and take care in choosing one useful for the problem (Sundararajan and Najmi, 2019). In addition, attributions are communicated at the level of input features, this entails some loss of information. Furthermore, attributions do not summarize the entire model behavior: this is closely tied with the principle of locality and limits tremendously the ability to explain models globally, rather than locally.

3 TOWARDS A TRUE XAI APPROACH

True Explainable AI system would allow to understand and validate how the AI system arrived at its conclusions.

As discussed above, Model induction XAI starts always with a “black box” AI model (which causes problems associated with the inability to fully understand the model or augment it with user expertise) and tries to give a best guess on how a given decision is made. The analogy is that if your complex car malfunctions and it is plugged to a diagnostic computer, it comes with a code which is not conclusive on what is the exact problem and it might mean investigating manually many parts of the car before fixing the problem.

Hence, there is a need for what we can call “True” XAI solutions generating fully transparent models which could be easily read, analyzed and augmented by the sector stake holders. The model generation steps could be easily tracked back to the data and the models could be audited before deployment to eliminate any bias and to ensure safe model operations which includes humans in the loop. Such XAI decisions give exact reasons to why a given output was generated and the output can be easily tracked. Back to the car malfunctioning analogy, the “True” XAI generates a very efficient car (with performance and options similar to the other complex car) which the driver and the mechanic understands exactly how it works and if it malfunctions, it will tells exactly what is the component which failed and how to rectify it and why this problem happened.

From the above discussion, it seems that offering the user with IF-Then rules which include linguistic labels appears to be an approach which can facilitate the explainability of a model output with the ability to explain and analyse the generated model. One AI technique which employs IF-Then rules and linguistic labels is the Fuzzy Logic System (FLS). However, FLSs are not widely explored as an XAI technique (and not even taught in many AI courses) and they donot appear in the analysis shown in Figure 1. One reason might be is that FLSs are associated with control problems and they are not widely perceived as a machine learning tool as they need the help of other techniques to learn their own parameters from data.

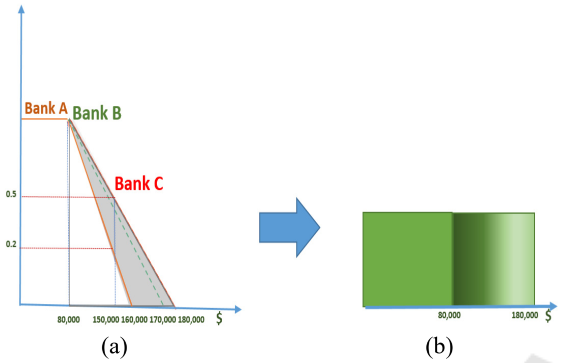


Figure 2: a) A Type-2 fuzzy set embedding the type-1 fuzzy sets for the linguistic label "*Low Income*" from experts in three banks. b) A graphical simplification of the type-2 fuzzy set in Figure 2a (Hagras, 2018).

Fuzzy Logic can model and represent imprecise and uncertain linguistic human concepts such as *Low*, *Medium*, *High*, etc. For example if a group of people were asked about the values they would associate with the linguistic concepts "*Low*" and "*High*" annual income and if Boolean logic was employed then we would have to choose a threshold above which income values would be considered "*High*" and below which they would be considered "*Low*". The first problem encountered is to identify a threshold that most people would agree on and this will be a problem as everyone has different idea what this linguistic label constitute.

On the other hand the linguistic labels "*Low*" and "*High*" could be represented by employing the type-1 fuzzy sets. In this representation, no sharp boundary would exist between sets and each value in the *x* axis can belong to more than one fuzzy set with different membership values. For example using type-1 fuzzy logic, $150,000 can belong to the "*Low*" and "*High*" sets but to different degrees where its membership value to "*Low*" could be 0.3 and to "*High*" is 0.7. This can mean that if 10 people were asked if $150,000 is *Low* or *High* income, 7 out of 10 would say "*High*", (i.e. membership value of 7/10=0.7) and 3 out of 10 would say "*Low*", (i.e. membership value of 3/10=0.3). Hence, fuzzy sets provide a means of calculating intermediate values between absolute true and absolute false with resulting values ranging between 0.0 and 1.0. Thus, fuzzy logic allows the calculation of the shades of grey between true/false. In addition, the smooth transition between the fuzzy sets will give a good decision response when facing the noise and uncertainties. Furthermore, FLSs employ linguistic IF-THEN rules which enable to represent the information in a human readable form which could be easily read, interpreted and analysed by the lay user.

As discussed in (Hagras, 2018), (Mendel, 2016), (Ruiz et al., 2019), (Sarabakha et al., 2017) the type-1 fuzzy sets are crisp and precise; hence they can handle only the slight uncertainties. However, different concepts mean different things to different people and in different circumstances. So assume as shown in Figure 2a, we asked three financial experts in three different banks (Bank A, Bank B and Bank C) to cast their opinions about what are the suggested ranges for "*Low*" income. As can be seen in Figure 2, each expert might come with different type-1 fuzzy set to represent the "*Low*" linguistic label. Another way to represent linguistic labels is by employing type-2 fuzzy sets as shown in Figure 2a which embeds all the type-1 fuzzy sets for Bank A, Bank B and Bank C within the Footprint of Uncertainty (FoU) of the type-2 fuzzy set (shaded in grey in Figure 2a). Hence, a type-2 fuzzy set is characterized by a fuzzy membership function, i.e. the membership value for each element of this set is a fuzzy set in [0,1], unlike a type-1 fuzzy set where the membership value is a crisp number in [0,1]. The membership functions of type-2 fuzzy sets are three dimensional and include a Footprint Of Uncertainty (FOU), this provide additional degrees of freedom that can make it possible to directly model and handle the uncertainties. More information about type-2 fuzzy sets and systems can be found in (Ruiz et al., 2019), (Sarabakha et al., 2017).

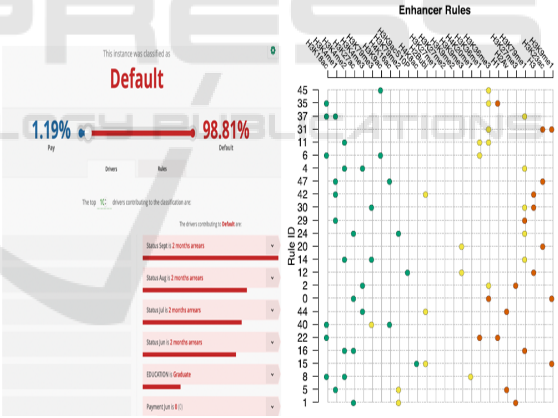
One misconception about type-2 fuzzy sets is that they are difficult to understand by the lay person. However, this is not the case as if experts are questioned about how to quantify a linguistic label, they will be sure about a core value (which has a common consensus across all experts), however they will struggle to give exact points of the boundaries of this linguistic label and there will uncertainty about the end points of a given linguistic label. Hence, a simplified version of a type-2 fuzzy set can be shown in Figure 2b where for the linguistic label "*Low*" income, there is a core value (shaded in solid green) of less than $80,000 which all experts agrees on and there is grey area (of shades of green) which goes between $80,000 to $180,000 of decreasing membership where there is uncertainty about the end points of the linguistic label where points beyond $180,000 are not recognised as "*Low*" income anymore (Hagras, 2018).

Another misconception of FLSs in general is that they are control mechanisms. This is not true as the area of Fuzzy Rule-Based Systems (FRBSs) generated from data has been active for more than 25 years. However, this was hindered by the FLSs incapability to handle systems with big number of inputs due to the phenomena known as curse of dimensionality where the FLS can generate long rules and huge rule bases which turn them to black boxes which are not easy to understand or analyse. Furthermore, FRBSs werenot able to handle easily imbalanced and skewed data (such as those present in fraud, bank default data, etc). However, recent work such as (Antonelli et al., 2017), (Sanz et al., 2015) was able to use evolutionary systems to generate FRBSs with short IF-Then rules and small number of rules in the rule base while maximizing the prediction accuracy. As this created sparse rule base not covering the whole search space, they presented a similarity technique to classify the incoming examples even if they do not match any fuzzy rule in the generated rule base. To do so, the similarity among the uncovered example and the rules was considered. They also presented multi-objective evolutionary optimization which was able to increase the interpretability (by reducing the length of each rule to include between 3 and 6 antecedents even if the system had thousands of inputs as well as having a small rule base) and maximize the accuracy of the FLS prediction. It was shown in (Antonelli et al., 2017), (Sanz et al., 2015) that such highly interpretable systems outperform decision trees like C4.5 by a big margin in accuracy while being easy to understand and analyze than the decision trees counterparts.

What is most important is that unlike other white box techniques, the FRBS generates IF-Then rules using linguistic labels (which can better handle the uncertainty in information) where for example in a bank lending application a rule might be: IF Income is *High* and *Home Owner* and Time in Address is *High* Then *Good* Customer. Such rule can be read by any user or analyst. What is more important is that such rules get the data to speak the same language as humans. This allows humans to easily analyze and interpret the generated models and most importantly augment such rule bases with rules which capture their expertise and might cover gaps in the data (for example, human experience can augment such historically generated rules with the human expertise to cover situations which did not happen before). This allows the user to have full trust in the generated model and also cover all the XAI components mentioned in (Gunning, 2017) related to Transparency, Causality, Bias, Fairness and Safety. Unlike the anchor rules mentioned in (Ribeiro, 2016b), humans do not make their decisions based on one single rule, they usually have Pros and Cons linguistic rules which humans balance and weigh in their mind and take a decision accordingly.

The second criteria which a type-2 XAI true explainable AI model provides is the ability to generate transparent outputs which could be easily understood and analysed by the lay user (as shown in Figure 3a for predicting credit cards defaults). Also it will be easy to follow the data route for the final decisions and its reasons.

Hence, viewing Figure 1, it can be seen that type-2 FLS and FRBSs can be best in explainability while striking a good balance to prediction accuracy when compared to other black box techniques. Furthermore, the type-2 FLSs could be used to explain the decisions achieved from more complex black box modelling techniques. Hence, the type-2 FLS and FRBSs can offer a very good way forward to achieve XAI which can be understood, analysed and augmented by the lay user.


This instance was classified as

Default

The top 10 drivers contributing to the classification are:

The drivers contributing to Default are:

Status Sept is 2 months arrears

Status Aug is 2 months arrears

Status Jul is 2 months arrears

Status Jun is 2 months arrears

EDUCATION is Graduate

Payment Jun is 0 (0)

(a) (b)

Figure 3: a)Fuzzy Logic instance drivers for example of 'Default' classification in Credit Card model (Adams and Hagras, 2020). b) Rules explaining enhancer and non-enhancer classification extracted from the XAI model in (Wolfe et al., 2021). Individual rules are horizontal lines on the plot and include up to three epigenetic marks per rule. The colour code represents classification of an epigenetic mark as high (green), medium (orange), or low (red).

4 REAL WORLD TRUE XAI DEPLOYMENTS

There has been several True XAI deployments in several real-world applications. For example, in (Wolfe et al., 2021), Type-2 Fuzzy based XAI was employed to predict the location of known enhancers with a high degree of accuracy. Enhancer malfunction is a key process that drives the aberrant regulation of oncogenes in cancer. Enhancer variants contribute more than any other known mechanism to heritable cancer predisposition. Enhancers are non-coding regions of the genome that control the activity of target genes. Recent efforts to identify active enhancers experimentally and *in silico* have proven effective. While these tools can predict the locations of enhancers with a high degree of accuracy, the mechanisms underpinning the activity of enhancers are often unclear. True XAI techniques was applied in (Wolfe et al., 2021) and gave very good accuracy of prediction close to opaque box models but additionally the XAI model provided insight into the underlying combinatorial histone modifications code of enhancers. In addition, the XAI model identified a large set of putative enhancers that display the same epigenetic signature as enhancers identified experimentally. Figure 3b (Wolfe et al., 2021) shows the extracted rules from the XAI model which revealed for the first time the mechanisms underpinning the activity of enhancers. ***This can open the way to new way for early detection and cancer treatment***

In (Andreu-Perez, 2021), the type-2 fuzzy based XAI model was employed for the analysis and interpretation of Infant functional near infrared spectroscopy (fNIRS), data in Developmental Cognitive Neuroscience (DCN). In the last decades, non-invasive and portable neuroimaging techniques, such as (fNIRS), have allowed researchers to study the mechanisms underlying the functional cognitive development of the human brain, thus furthering the potential of DCN. However, the traditional paradigms used for the analysis of infant fNIRS data are still quite limited. In (Andreu-Perez, 2021), they introduced a Multivariate Pattern Analysis for fNIRS data, xMVPA, that was powered by true (XAI). The proposed approach was exemplified in a study that investigates visual and auditory processing in six-month-old infants. xMVPA not only identified patterns of cortical interactions, which confirmed the existent literature; in the form of conceptual linguistic representations, it also provided evidence for brain networks engaged in the processing of visual and auditory stimuli that were previously overlooked by other methods, while demonstrating similar statistical performance. The XAI model did show very important results in that the model for the developing brain has similar modules and interconnections as the adult neural system for face perception presented by (Haxby et al., 2014) suggesting that by 6 months of age the cortical activity associated with face processing is already similar to that of mature brains. However, the model revealed that the inter-regional interactions between the temporal and prefrontal cortex might be specific to speech-like sounds. This XAI model also showed a selective pattern of activation over the temporal cortex that is specific to visual vs. auditory stimuli (Andreu-Perez, 2021). Specifically, the channels of the temporal cortex which are active in response to the visual stimulus are instead inactive in response to the auditory stimulus. This confirmed the multifaceted role of temporal cortex in the processing of sensory stimuli thereby some areas are dedicated to visual processing whilst others are associated with auditory processing (as reported by (Nolan and Altman, 2001)). The temporal cortex will form the core system for processing non-speech auditory stimuli, while the prefrontal cortex will form the extended system for processing the emotion associated with the auditory stimulus. When inactive, the occipital cortex enables the occurrence of these patterns. Hence, the work in (Andreu-Perez, 2021) revealed new brain regions activation and interactions not yet established for the developing brain (as shown in Figure 4). Learning new cortical pathways directly from the neuroimaging data is of fundamental significance in DCN research to shed light on functional brain development in absence of established assumptions. ***Hence, this True XAI model can open the way to understand the development of the human brain and this can allow the early detection and management of atypical functional brain development like Autism.*** Autism early detection and intervention can divert individuals from sustained care pathways that, beyond being expensive they damage the quality of life of individuals. Furthermore, this can help in the early intervention and communications with practitioners and parents, to avoid social stigmas and enable professional support as early as possible where early intervention can remarkably improve these children's functional and social skills, improving their capacities to avoid later dependency on the state social system, and fending for themselves during adulthood.

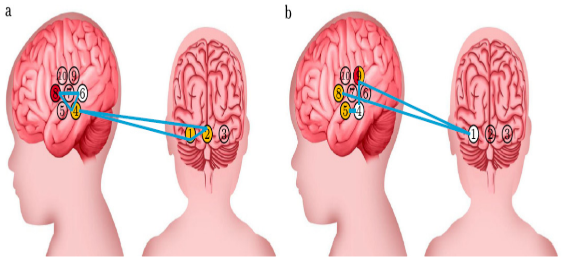


Figure 4: Patterns of cortical networks delineated by xMVPA. The patterns (cyan) identified by the xMVPA delineate the contributions between brain regions evoked by a visual and b auditory stimuli. The colour of the channels denotes their level of activity: inactive (white), active (amber), and very active (red), and uncoloured for channels that do not belong to any pattern (Andreu-Perez, 2021).

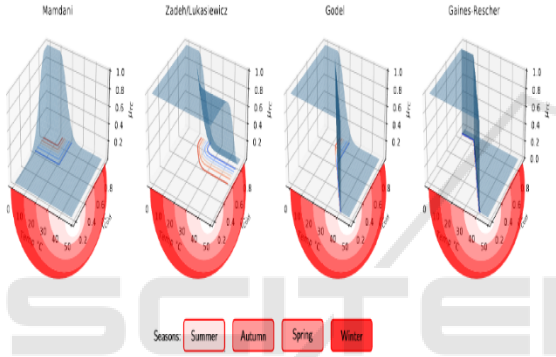
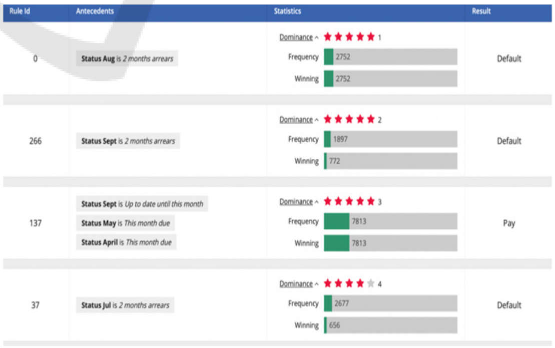


Figure 5: A comparison of Temporal Type-2 Fuzzy Set for the conceptual label (CoL) 'Cold' for feature thermal concept constructed with the most commonly used fuzzy relations namely Mamdani, Zadeh/Lukasiewicz, Godel, and Gaines-Rescher (Kiani et al, 2022).

The work in (Kiani et al., 2022), enabled XAI models to be handle time dependent applications To account for the temporal component, where they presented *Temporal Type-2 Fuzzy System Based Approach* for time dependent XAI systems (TXAI), which can account for the likelihood of a measurement's occurrence in the time domain using (the measurement's) frequency of occurrence. In Temporal Type-2 Fuzzy Sets (TT2FSs), a four dimensional(4D) time-dependent membership function (as shown in Figure 5) is developed where relations are used to construct the inter-relations between the elements of universe of discourse and its frequency of occurrence. TXAI can also outline the most likely time dependent trajectories using the frequency of occurrence values embedded in the TXAI model; viz. given a rule on a determined time, what will be the next most likely rule at a subsequent time point. In this regard, this TXAI ***system can have profound implications for delineating real-life time-dependent processes, such as behavioural or biological modelling across time***.

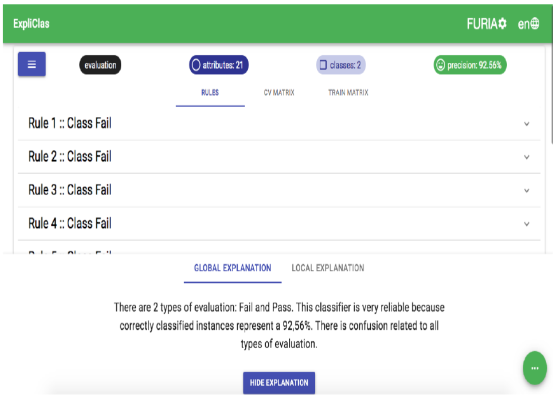
In (Adam and Hagras, 2020), a true XAI approach was presented to develop risk management framework for the implementation of AI in banking with consideration of explainability to enable AI to achieve positive outcomes for financial institutions and the customers, markets and societies they serve. This work showed that the type-2 based true XAI model delivered very good performance which is comparable to or lagging marginally behind the Neural Network models in terms of accuracy, but outperform all models for explainability, thus they are recommended as a suitable machine learning approach for use cases in financial services from an explainability perspective. This research is important for several reasons: (i) there is limited knowledge and understanding of the potential for Type-2 Fuzzy Logic as a highly adaptable, high performing, explainable AI technique; (ii) there is limited cross discipline understanding between financial services and AI expertise and this work aims to bridge that gap; (iii) regulatory thinking is evolving with limited guidance worldwide and this work aims to support that thinking; (iv) it is important that banks retain customer trust and maintain market stability as adoption of AI increases. Figure 6 shows the global rules extracted from data via the XAI model for Credit Card Defaulting prediction case. This shows the ability to generate from data rules which could be easily understood, analysed and augmented by the business user which are very important factors ***for the wide deployment and acceptance of AI models in the finance sector.***



Rule Id	Antecedents	Statistics	Result
0	**Status Aug** is *2 months arrears*	Dominance ★★★★★ 1 Frequency 2752 Winning 2752	Default
266	**Status Sept** is *2 months arrears*	Dominance ★★★★★ 2 Frequency 1897 Winning 772	Default
137	**Status Sept** is *Up to date until this month* **Status May** is *This month due* **Status April** is *This month due*	Dominance ★★★★★ 3 Frequency 7813 Winning 7813	Pay
37	**Status Jul** is *2 months arrears*	Dominance ★★★★☆ 4 Frequency 2677 Winning 656	Default

Figure 6: Top global rules for Credit Card Default use case (Adams and Hagras, 2020).

In (Alonso and Casalino, 2019), they presented an XAI tool, called ExpliClas, with the aim of facilitating data analysis in the context of the decision making processes to be carried out by all the stakeholders involved in the educational process.



Rule 1 :: Class Fail

Rule 2 :: Class Fail

Rule 3 :: Class Fail

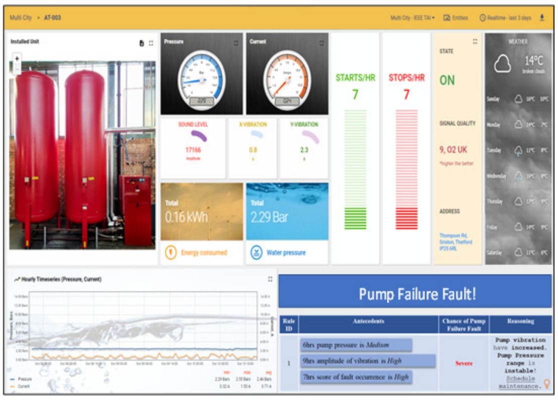
Rule 4 :: Class Fail

There are 2 types of evaluation: Fail and Pass. This classifier is very reliable because correctly classified instances represent a 92,56%. There is confusion related to all types of evaluation.

Figure 7: Example of global explanation obtained with ExpliClas (Alonso and Casalino, 2019).

ExpliClas provided illustrative examples of both global (shown in Figure 7) and local explanations related to the given dataset. In addition, ExpliClas automatically generated multimodal explanations which consisted of a mixture of graphs and text. These explanations look like natural, expressive and effective, similar to those expected to be made by humans. It is worth noting that the rationale behind ExpliClas is completely transparent to the user, which can understand the reasoning that leads to a given output

In (Upasane et al., 2023), they presented a type-2 fuzzy-based Explainable AI (XAI) system for predictive maintenance within the water pumping industry (as shown in Figure 8). The proposed system is optimised via Big-Bang Big-Crunch (BB-BC), which maximises the model accuracy for predicting faults while maximising model interpretability. They evaluated the proposed system on water pumps using real-time data obtained by their hardware placed at real-world locations around the United Kingdom and compared their model with Type-1 Fuzzy Logic System (T1FLS), a Multi-Layer Perceptron (MLP) Neural Network, deep neural networks learning method known and decision trees (DT). The proposed system predicted water pumping equipment failures with good accuracy (outperforming the T1FLS accuracy by 8.9% and DT by 529.2% while providing comparable results to SAEs and MLPs) and interpretability. The system predictions comprehend why a specific problem may occur, which leads to better and more informed customer visits to reduce equipment failure disturbances. It was shown that 80.3% of water industry specialists strongly agree with the model's explanation, determining its acceptance. ***This will allow to the wide deployment of XAI within the predictive maintenance industries.***



Rule ID	Antecedents	Chance of Pump Failure Fault	Reasoning
1	6hrs pump pressure is *Medium* 9hrs amplitude of vibration is *High* 7hrs score of fault occurrence is *High*	Severe	Pump vibration have increased. Pump Pressure range is instable! Schedule maintenance.

Figure 8: Individual telemetry unit's data with fault's explanation (Upasane et al., 2023).

5 CONCLUSIONS AND FUTURE WORK

This paper presented the notion of "True" XAI models which can be easily analysed, understood and augmented by the sector stake holders. Such XAI models outputs can be easily analysed and understood by the lay users. We have shown that such "True" XAI models can lead to major breakthrough in various sectors and lead to major discoveries and innovations. Most importantly such XAI systems can lead to the wide deployment and trust of AI in various domains where human trust is needed and heavy regulations are in place. From my point of view, such "True" XAI systems will be very important for the safe, secure and fair applications of AI. Hence, I expect the growth of such "True" XAI systems in various applications all over the World.

REFERENCES

Adams, J., Hagras, H. (2020). A Type-2 Fuzzy Logic Approach to Explainable AI for regulatory compliance, fair customer outcomes and market stability in the Global Financial Sector, *Proceedings of the 2020 IEEE International Conference on Fuzzy Systems, Glasgow, UK, July 2020.*

Alonso, J.M., Casalino, G. (2019). Explainable Artificial Intelligence for Human-Centric Data Analysis in Virtual Learning Environments. *In: Burgos, D., et al. Higher Education Learning Methodologies and Technologies Online. HELMeTO 2019. Communications in Computer and Information Science*, Vol.1091. Springer, Cham

Andreu-Perez, J., Emberson, L., Kiani, M., Filippetti, M., Hagras, H., Rigato, S. (2021). Explainable artificial intelligence based analysis for interpreting infant fNIRS data in developmental cognitive neuroscience, *Communications Biology*, Vol. 4, No. 1, pp. 1-13, September 2021.

Antonelli, M., Bernardo, D., Hagras, H., Marcelloni, F. (2017). Multi-Objective Evolutionary Optimization of Type-2 Fuzzy Rule-based Systems for Financial Data Classification. *IEEE Transactions on Fuzzy Systems*, Vol. 25, No. 2, pp. 249-264, April 2017.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., Samek, W (2015). On pixelwise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PLoS ONE*, Vol. 10, No. 7, 2015.

Chavatte, L (2018). Artificial Intelligence in Europe Report: At a glance. Microsoft Pulse. https://pulse. microsoft.com/en/transform-en/na/fa1-articial-intelligence-report-at-a-glance/#:~:text=65%25%20is%20expecting%20AI%20to,key%20topic%20for%20executive%20management.

Gunning, D, (2017). Explainable Artificial Intelligence. http://www. darpa.mil/program/explainable-articial-intelligence, 2017.

Hagras, H. (2018). Towards Human Understandable Explainable AI", *IEEE Computers*, Vol.51, No.9, pp. 28-26, September 2018.

Haxby, J. V., Connolly, A. C. & Guntupalli, J. S (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience,* Vol. 37, pp. 435-456, 2014.

Kiani, M., Andreu-Perez, J., Hagras, H (2022). A Temporal Type-2 Fuzzy System for Time-dependent Explainable Artificial Intelligence, *IEEE Transactions on Artificial Intelligence*, pp.1-15, September 2022.

Montavon, G., Samek, W., Müller,K. (2018) Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, Vol.73, pp.1-15, 2018.

Mendel, J., Hagras, H., Sola, H., Herrera, F. (2016) Comments on: Interval Type-2 Fuzzy Sets are generalization of Interval-Valued Fuzzy Sets: Towards a Wider view on their relationship. *IEEE Transactions on Fuzzy Systems,* Vol.24, No.1, pp.249-250, February 2016.

Nolan R. Altman, B. (2001). Brain Activation in Sedated Children: Auditory and Visual Functional MR Imaging. *Pediatric Imaging, Vol.1* 221, pp.56-63, 2001.

Press, G. (2019), AI And Automation 2019 Predictions From Forrester, https://www.forbes.com/sites/gilpress/2018/11/06/ai-and-automation-2019-predictions-from-forrester/#d46ec454cb57.

Ribeiro, M., Singh, S., Guestrin, C. (2016a) why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 2016 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016

Ribeiro, M., Singh, S., Guestrin, C. (2016b) Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance. *ArXiv e-prints*, November 2016.

Ruiz, G., Hagras, H., Pomares, H., Rojas, I. (2019) Towards a Fuzzy Logic System Based on General Forms of Interval Type-2 Fuzzy Sets, *IEEE Transactions on Fuzzy Systems*, Vol. 27, No.12, pp. 2381-2396, February 2019.

Sanz, J., Bernardo, D., Herrera, F., Bustince, H., Hagras, H., (2015). A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications with Imbalanced Data. *IEEE Transactions on Fuzzy Systems*, Vol.23m No.4, pp.973-990, August 2015.

Sarabakha, A., Imanberdiyev, N., Kayacan, E., Khanesar, M., Hagras, H. (2017). Novel Levenberg–Marquardt based learning algorithm for unmanned aerial vehicles" *Journal of Information Sciences*, Vo.417. pp. 361-380, November 2017.

Sundararajan, M., Nahmi, A., 2019. The many Shapley values for model explanation. *Proceedings of Machine Learning Research*, Vol.119, pp. 9269-9278, 2019.

Upasane, S., Hagras, H., Anisi, M., Savill, S., Taylor,I., Manousakis, K. (2023). A Type-2 Fuzzy Based Explainable AI System for Predictive Maintenance within the Water Pumping Industry. *IEEE Transactions on Artificial Intelligence*, May 2023.

Wolfe, J., Mikheeva, L., Hagras, H., Zabet, N.(2021). An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila, *Genome Biology*, Vol. 22, No.1, pp.1-23, December 2021.