Embedding Transparency in Artificial Intelligence Machine Learning Models: Managerial Implications on Predicting and Explaining Employee Turnover

Soumyadeb Chowdhury[1], Sian Joel-Edgar[2], Prasanta Dey[3],**
Sudeshna Bhattacharya[4], Alexander Kharlamov[3],
Operations, Information and Management Science[1], Operations and Information Management Department[2], Work and Organisation Department[3]

Toulouse Business School (France)[1]; New College of the Humanities, Northeastern University (UK)[2], Aston Business School, Aston University (UK)[3,4]

**** Corresponding Author: Soumyadeb Chowdhury (S.Chowdhury@tbs-education.fr)**

Abstract

Employee turnover (ET) is a major issue faced by firms in all business sectors. Artificial intelligence (AI) machine learning (ML) prediction models can help to classify the likelihood of employees voluntarily departing from employment using historical employee datasets. However, these AI-based ML models lack transparency, making it difficult for HR managers to understand the rationale behind the AI predictions. If managers do not understand how and why outputs are generated by AI, it is unlikely to augment data-driven decision-making and bring value to the organisations. The main purpose of this article is to demonstrate the Local Interpretable Model-Agnostic Explanations (LIME) software package which can qualitatively and intuitively explain the predictions generated by AI-based ML models to HR managers. From a theoretical perspective, we contribute to the International Human Resource Management literature by presenting a conceptual review of AI algorithmic transparency and then discussing its significance to sustain competitive advantage by using the principles of resource-based view theory. We also offer a transparent AI-based implementation framework using LIME which will provide a useful guide for HR managers to increase the explainability of the AI-based ML models, for mitigating trust issues in data-driven decision-making.

Keywords

Artificial Intelligence, Machine Learning, Employee Turnover, AI Transparency, Local Interpretation, Model Explainability, Human Intelligence.

Introduction

Employee turnover (ET) is defined as an employee leaving an organisation and the termination of the contract, both formal and psychological (Shaw *et al.*, 2005). ET is important for organisations due to a variety of reasons: i) ET is expensive (O'Connell and Kung, 2007), it costs time, money and other organisational resources; ii) New talent acquisition is often very challenging (Ju and Li, 2019); and iii) ET impacts business performance at multiple levels (Shaw *et al.*, 2005). As a result, ET has been used as an indicator of organisational effectiveness (Edgar et al, 2017) and it is one of the main metrics used in human resource management (Allen, 2008). A multitude of factors impact on an employee's decision to exit a job role within an organisation (Mishra and Sahoo, 2018), making it particularly challenging for a line manager to predict such a decision or take steps to mitigate the potential turnover. Furthermore, for line managers within organisations faced with the task to identify employees at risk of leaving, accessible decision-making tools are either absent or inaccessible at best (Huselid, 2018). This is where Artificial intelligence (AI) provides avenues for developing tools to facilitate strategic HR decision-making (Johnson et al. 2020).

AI refers to a set of techniques and algorithms that can automatically integrate, process and learn from data, and apply those learnings to achieve specific objectives and tasks (Haenlein and Kaplan, 2019). AI-based techniques can assist in predicting staff turnover, i.e., likelihood of an employee leaving the organisation. However, these techniques currently lack transparency and explainability, which makes it difficult for managers to trust the output of AI (Agarrwal et al., 2020; Bieda et al., 2020). In the context of transparency, it is a form of ideal that should facilitate revealing how data is integrated into an algorithm, processed by the algorithm and the knowledge that is gained using that data (Cheng and Hackett, 2019). The issue with explainability is that business managers do not know how AI-based machine learning (ML) algorithms generate the outputs by processing the input data because the algorithm is either proprietary or that the mathematical computational models used in the algorithm are very complex to understand (Shin and Park, 2019). In this context, existing literature has discussed that the time, effort, and resources invested by the organisations in AI systems has not translated into business value and productivity in many of the cases (Fountaine et al., 2019). Limited transparency and explainability of output responses generated by the AI systems has emerged as a key barrier to experiencing anticipated benefits by confidently turning data-centric decisions into effective actionable strategies (Shin et al., 2019; Makarius et al., 2020).

The primary goal of embedding transparency within AI-based ML models is to help the decision-making authorities understand what the AI system is doing, how it is generating the output responses and why a particular response is generated (Choudhury et al., 2020). This will help these business users to confidently assess the accuracy of the responses based on their own tacit domain expertise, which will increase trust in these systems (Cowgill and Tucker, 2020). The ability to get explanations for the output responses will also reduce biases in business processes, operations, and decision-making, thus enhancing fairness (Satell and Sutton, 2019). For example, gender discrimination in hiring employees and setting up credit card borrowing limits stemming from AI systems has led to mistrust among both businesses and their consumers, demonstrating the need for AI transparency (BBC, 2019). Furthermore, AI transparency can also aid in identifying and resolving flaws within ML models stemming from improper training datasets (input issues), wrong settings, configurations and hyperparameters (algorithmic issues), and overfitting or underfitting models, which will enhance the value offered by these AI-based systems.

With regards to enhancing the transparency of AI-based ML models, *Local Interpretable Model-Agnostic Explanations* (LIME) is a software package that can explain the predictions made by both linear and non-linear ML algorithms. LIME presents easy to interpret textual and visual representations that can provide a qualitative understanding of the relationships between the ML output responses and input variables (Ribiero et al., 2016). Therefore, the software explains how and why certain decisions were made by the AI algorithm (Zhang et al., 2019), unearthing the opaqueness of these algorithms to the users. The main purpose of this article is to demonstrate an implementation framework using LIME that will help HR decision-makers to understand the logic behind the decisions made by AI-based machine learning models.

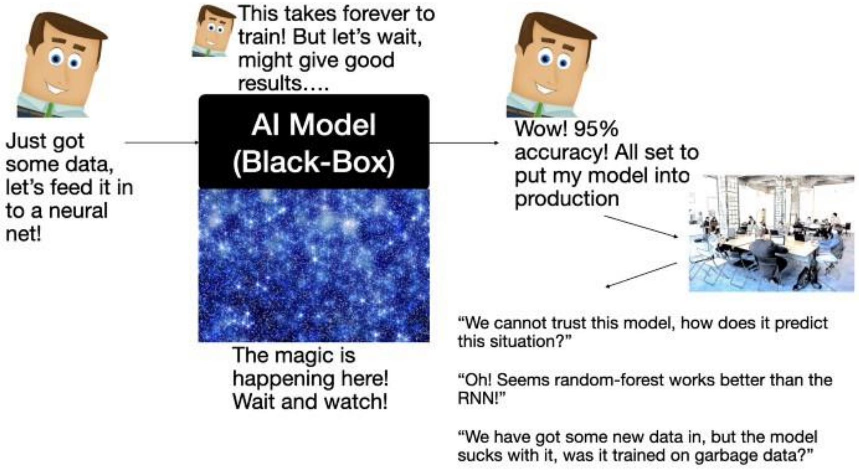

This takes forever to train! But let's wait, might give good results….
Just got some data, let's feed it in to a neural net!
AI Model (Black-Box)
Wow! 95% accuracy! All set to put my model into production
The magic is happening here! Wait and watch!
"We cannot trust this model, how does it predict this situation?"
"Oh! Seems random-forest works better than the RNN!"
"We have got some new data in, but the model sucks with it, was it trained on garbage data?"

Figure 1: AI Transparency problem (Chatterjee, 2020)

Recent studies have emphasised the benefits of employing AI-based ML tools in HR processes (Daugherty et al., 2019), applications of AI in talent acquisition (Gusnadi and Hermawan, 2020), managerial implications pertaining to AI implementation (Morse, 2020; Suen et al., 2019), the impact of deploying these tools on the job roles, responsibilities, tasks and meaningfulness of work (Iansiti and Lakhani, 2020; Wilson et al., 2017), and the importance of embedding transparency within the AI algorithms (Chowdhury, 2020; Glikson and Woolley, 2020). Currently, how transparency can be achieved, and which organisational resources are required to unlock the potential of AI transparency is under-researched within the literature (Amabile, 2020; Makarius et al., 2020). Therefore, the work reported in this paper will aim to bridge this knowledge gap (also shown in Figure 1) drawing from the resource-based view theory (RBV) of the firm. RBV is one of the most widely applied theoretical perspectives to explain how resources within an organisation can help enhance business performance and competitiveness (Barney, 2001). The existing literature has also demonstrated appropriateness of RBV to be applied as a theoretical lens for developing distinctive and hard-to-imitate capabilities (such as AI transparency) in a turbulent and technology-driven business environment (Bromiley and Du, 2016; Mikalef and Gupta, 2021).

Building on the previous literature concerning AI in HRM decision-making and the significance of enhancing the transparency and explainability of AI algorithms leads to the following two research questions that motivate our current research:

- *RQ1: How can we employ AI-based machine learning algorithms to predict staff turnover?*
- *RQ2: How can we embed transparency in machine learning algorithms to explain the rationale for the output generated by these algorithms to predict employee turnover?*

The answers to these research questions will enhance HR business managers' ability to confidently use AI-based decision support systems employing ML algorithms (RQ1) and understand the output responses predicted by these algorithms, for developing strategies and initiatives towards staff retention and talent management (RQ2).

Answering these questions is important as organisational scholars have indicated and acknowledged the increasing use and impact of AI in HR decision-making processes for gaining competitive advantage (Gunasekaran et al., 2017). The significance of embedding transparency in AI-based decision support systems, i.e., how the outputs are generated by AI algorithms is well articulated and clear (Chowdhury, 2020), which is also less developed in the general management and IHRM academic literature. Therefore, with advances in technological and algorithmic innovations, the problem has shifted from collecting huge volumes of data, turning data into knowledge and conclusions (Kersting and Meyer, 2018), to understanding how AI algorithms generate these conclusions (Gulliford and Dixon, 2019). This will facilitate building trust among the managers and turn these conclusions into actionable insights (developing employee retention strategies based on the evidence drawn from the data about employees).

We integrate theoretical tenets of RBV with AI literature to frame the implications of this work and contributions to the IHRM field in the following ways. We examine the explicability and transparency in AI-based ML Models as a strategic resource considering the RBV theory (Boxall, 1996) in the context of ET. Resources can be both tangible and intangible assets associated with the firm (Caves, 1980). In this context, transparency within AI systems is intangible because it represents the system's quality and characteristics (Haibe-Kains et al., 2020). From a firm's perspective, technology is often one of its core strategic resources, which is essential to gain and sustain a competitive advantage (Alalie et al. 2018). The effectiveness of a technological resource greatly depends on its adoption and use (Wernerfelt, 1984). Lack of trust in technologies will negatively impact its adoption and subsequent use to generate value (Andriopoulos & Lewis, 2009). Therefore, for the AI-based ML models to become an effective strategic resource, they must incorporate transparency (Raisch & Krakowski, 2021), so that decision-makers can leverage the analytical capabilities of AI to augment and drive data-centric decision-making.

We contribute to the practice of AI by developing an implementation framework that embeds transparency in AI-based decision systems. This will assist the HR managers to understand the decisions predicted by the AI algorithms and the mechanics behind them. Transparency and explainability will be essential for the current and future adoption of AI applications, in the vein of augmenting and assisting human intelligence in the decision-making process. This will lead to advocating fair and responsible use of AI within organisations (Shrestha et al., 2020; Brock et al., 2019).

The paper is structured as follows. The next section provides an overview of the ET literature, AI algorithms and transparency, followed by the methodology and implementation framework proposed to predict ET using machine learning and embedding LIME. Next, the results and sensemaking techniques used to interpret the results of explainable AI are presented followed by the discussion on the role of the symbiotic relationship between human intelligence (HI) and AI in decision-making. Finally, conclusions are presented with a set of agendas shaping the future direction of research in this emerging and unexplored domain of embedding transparency in AI decision-making.

Literature Review

This section provides an overview of the literature pertaining to employee turnover, AI algorithms and AI transparency. The section concludes by outlining the knowledge gap in the literature in the context of employing AI to aid data-driven decision-making and embedding transparency into AI algorithms.

Employee Turnover

Human resources literature over the 20th century within post-industrial societies has attempted to unlock some of the factors underlying employee turnover with considerable success such as psychological contract breach (Shaw et al., 2005; Robinson and Morrison, 2000), job satisfaction, commitment and trust (Timming, 2012; Farrington, 2008; Eisenberger et al., 2002; Eisenberg, Fasolo and Davis-LaMastro, 1990). Further, researchers have been able to identify job characteristics that link to motivation and eventual retention of employees within the organisation (Hackman and Oldham, 1974; Ali et al., 2014). In the context of knowledge workers, situational factors such as autonomy, feedback, skill variety and task significance are particularly salient for important employee outcomes such as job satisfaction and motivation, which are significant pre-cursors to turnover intentions (Degbey et al., 2020; Neeley, 2017; Cornell and Shapiro, 1987). The Job Characteristics Model (JCM) puts forward five key characteristics that shape employees' sense of motivation for a job: skill variety, task identity, task significance, autonomy and feedback (Hackman and Oldham 1975). These characteristics in their view increase the meaningfulness of the work, in the context of knowledge work thus leading to motivation, job satisfaction and reduced chances of turnover (Mishra and Sahoo, 2018; Sears et al., 2013; Hackman and Oldham, 1975).

AI Classification and Algorithms

Research evidence suggests that AI is increasingly used for HR processes and tasks, such as selecting applicants for jobs and scheduling logistics (von Krogh, 2018). Machine learning and other AI technologies have been particularly used to develop capacity, for example, a hospital in Boston successfully forecast COVID-19- related clinical demands during the ongoing pandemic crisis (Stevens et al., 2020). AI-enabled systems have the capability to process big data (characterised by 5Vs – Volume, Velocity, Variability, Variety and Veracity) generated in human resource processes to automatically provide valuable insights to a decision-maker (workforce analytics – Makarius et al., 2020; Huselid, 2018), increase automation of routine, repetitive and trivial tasks (digital assistants - Johnson et al., 2020; Gusnadi and Hermawan, 2020), enable HR department in recruiting and improving candidate experience (recommendation engine – Jumar et al., 2019; Suen et al., 2019), improve employee engagement and experience within the organisation, through collaborative and personalised learning (digital assistants and machine learning –Gusnadi and Hermwan, 2020; Tambe et al., 2019; Malone, 2018;). Appendix 1 presents an overview of five commonly used ML classification-based algorithms. Based on the type of learning involved, AI algorithms can be classified into three categories (Di Vaio et al., 2020; Haenlein and Kaplan, 2019; Cohen, 2019; Davenport and Bean, 2017):

- *Supervised learning:* These methods map a given set of inputs to a labelled set of outputs, i.e. learn from the inputs and corresponding labelled output variables in the dataset, and then apply the learning for new cases introduced in the dataset. These algorithms are commonly used in solving classification and regression problems such as natural language processing (sentiment analysis), image recognition (e.g. face recognition) and financial forecasting (e.g. fraud detection).
- *Unsupervised learning:* These methods work with unlabelled data, i.e. the output variable is not labelled unlike supervised learning, therefore the algorithm needs to identify trends and patterns within the dataset and learn from these patterns. These algorithms find their use in segmentation and clustering problems such as trend detection in a weather dataset, customer segmentation for targeted marketing, and dimensionality reduction such as big data visualisation through variable consolidation and feature elicitation.

- *Reinforcement learning:* These algorithms aim to maximize the output by selecting the most appropriate decision from a set of input decisions, i.e. determine the optimal action for the most favourable outcome. They are primarily used in real-time decision-making in board and video games, robotics process automation such as robots in supply chain warehouse, digital personal assistants which adapt the responses based on previous experience for similar queries and self-driving cars.

This paper will employ supervised machine learning to predict the probability of an employee leaving the organisation based on a set of input variables and corresponding labelled output variables (i.e., employee attrition having a binary response YES or NO). Therefore, we are investigating a classification problem, where the algorithm will predict a class for each case (employee), which can be either YES or NO. Appendix-1 presents an overview of five key machine learning algorithms which are commonly used for classification in real-life applications and are also included within the machine learning libraries in popular open-source programming tools such as Python, R, JAVA and Go (Kotsiantis et al., 2007; Osisanwo et al., 2017; Soofi and Awan, 2017).

AI Transparency

In her book "Weapons of Math Destruction" Cathy O'Neil (2016) describes how AI algorithms, so abundant today, are 'opaque, unregulated, and uncontested, even when they're wrong'. In response to this known issue, there have been ethical guidelines posited which address transparency, justice and fairness, non-maleficence, responsibility and privacy (Glikson and Woolley, 2020). Three areas of particular focus to address the opaque and black box nature of AI algorithms are fairness, accountability and transparency (Shin and Park, 2019).

There are several reasons why transparency is particularly necessary in an HR context (Chowdhury et al., 2020; Silvernam, 2020). For example, in the employee recruitment process, if the outcome of an AI algorithm is unfavourable for an applicant, the applicant and HR managers (unless they have been trained), have no mechanism for discovering why the applicant was unsuccessful, and consequently, the applicant cannot knowingly improve his or her skill set. It is assumed in this argument that there is a way of controlling the input data and changing the outcome. This may not always be the case, as identified by Crain (2018), whereby transparency can be disconnected from power. This leads to the second area in which transparency is necessary, to address bias.

Certain groups have been found to be disproportionally disadvantaged in AI algorithms, e.g. black faces associated as gorillas (Dougherty, 2015) and Asian people categorised as blinking (Wade, 2010). If a proportion of society is consistently marginalised in the job market or in a particular organisation, HR managers need to answer user and societal questions. If users or HR managers do not understand the algorithms' affordances and variants, this can result in an inability to use the algorithms effectively to recruit and retain the best possible staff and to potentially be swayed by prejudice (Chowdhury et al., 2020; Shin and Park, 2019). It should not be acceptable that 'blame' for such inappropriate outcomes such as prejudice fall upon a 'mathematical model'. Ownership of the AI algorithm and its results may be placed on HR managers, and as such, they would need to know the rationale for the data input choices and results (Davenport and Ronanki, 2018).

The concept of transparency is a complex one and Ananny and Crawford (2018) outline many reasons why transparency is sometimes not needed, or the use of transparency is not straightforward. It is acknowledged, however, that a key to transparent AI systems is trust. Either users (including HR managers) understand and trust the algorithms, or they place trust in third parties who do understand the complexities of the algorithm (Glikon and Woolley, 2020). If HR managers understand the system, they then trust the designers and developers of the system and use the system effectively (Lee and Boynton, 2017).

The explainability of the AI algorithms will depend on several factors outlined below (Bieda, 2020; Pigni et al., 2016): *data characteristics*, which includes volume, heterogeneity, variability and velocity of the data; *number of input variables* used to train the dataset; *types of relationships* between the variables (linear or non-linear), *complexity of the algorithms* (i.e. whether these can be explained by mathematical functions or use neural network architecture having hidden layers between the input variables and output response); *type of learning employed*; *quality of the dataset* characterised by consistency, completeness, and minimal outliers. The existing research (Pillai and Sivathanu, 2020) in this area have reported that the ease to interpret and explain the response generated by AI algorithms both at global and local level decreases with the complexity of algorithms, increasing the size of the dataset and increasing the number of input variables as visualised in Figure 2.

In the context of explaining AI outputs, feature *importance plots* have been used to determine the key input variables in the dataset which may contribute to the output by assigning weight for each variable and visualizing them either using hierarchical trees or histogram representation (Liu et al., 2012). These plots do not provide information about the direction of the relationship, i.e. whether it is linear or non-linear, which requires additional effort to understand this dimension. Therefore, *partial-dependency plots* can be used by selecting 'n' number of highly weighted variables from feature importance plots, to identify the relationship between each of the input variables and the outcome variable in distinct plots for a given dataset (Krause et al., 2016).

There are three problems with partial-dependency plots (Jergensen et al., 2020): (1) It cannot be produced for advanced complex algorithms implementing non-linear classifiers such as neural networks (deep learning) because these algorithms use a network architecture having additional hidden layers between the input and output layer; (2) These plots can give the users an idea of the global behaviour of a machine learning algorithm, i.e. the features that are important in the global dataset, and make an assumption that this importance leads to the response produced by the algorithm. However, they fail to provide information explaining the output produced by the algorithm at a local level (i.e. for each specific case); (3) the number of variables and complexity of the relationship between these variables increases with the volume of the data, which will increase the number of plots, and therefore makes the process inefficient, cumbersome and difficult to interpret.

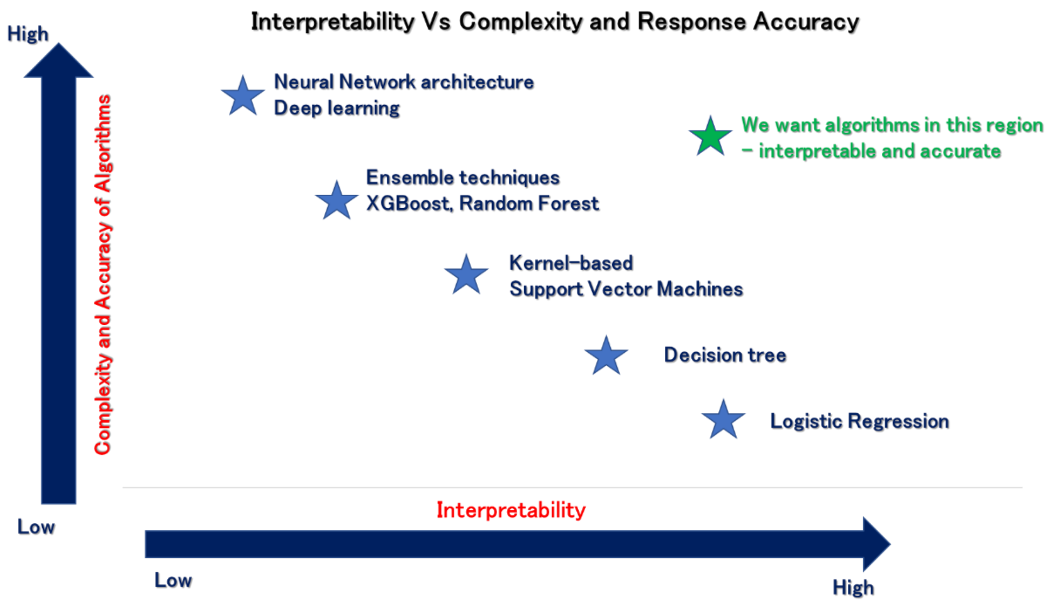

Interpretability Vs Complexity and Response Accuracy
High
Complexity and Accuracy of Algorithms
Low
Neural Network architecture
Deep learning
We want algorithms in this region
– interpretable and accurate
Ensemble techniques
XGBoost, Random Forest
Kernel-based
Support Vector Machines
Decision tree
Logistic Regression
Interpretability
Low
High

Figure 2: Comparing interpretability of AI classification algorithms

Knowledge Gap

The HRM literature has presented conceptual frameworks on applications of AI within the domain but does not demonstrate actual implementation of these frameworks through empirical studies using AI-based algorithms to understand the limitations posed by lack of transparency in these algorithms for the managers. Studies reporting AI frameworks embedding explainability is sparse not only in the HR literature but also in the business and management domain. This warrants further investigation in the emerging area of AI transparency to increase trust in automated analytical capabilities offered by emerging software technologies in the digital era, and supplement it with intuitive and social intelligence of managers based on their tacit experience to facilitate data-driven strategic decision-making. Increasing the transparency and explainability of AI algorithms' responses can help to enhance the trust of managers in these models because it provides valuable knowledge on the accuracy, relevance and process employed by those algorithms. Therefore, AI transparency has the potential to enable strategic change within the HR practices and policies not only at the global level (all staff within the organisation) but locally as well (per staff basis, i.e., catering to individual needs).

Methodology

The methodology employed in this paper draws from the business and information management literature used to develop a predictive analytics data science tool for qualitative researchers (Ciechanowski et al., 2020), principles of employing machine and deep learning algorithms (Shrestha et al., 2020), and machine learning models used for forecasting (Hwang et al., 2020). We present a ML implementation framework embedding transparency to predict employee turnover discussing the steps below.

Data Preparation and Pre-processing (Figure 3)

The ET data used for this study is a simulated dataset (Kaggle, 2020) created by IBM Watson based on real-life information. This has been used to test the accuracy of IBM Watson, and used by practitioners, therefore it is deemed suitable for this study. The dataset is pruned (i.e., cleaned) by eliminating: (1) rows with missing values; (2) variables (columns) that are inconsistent across the datasets to avoid data inconsistency.

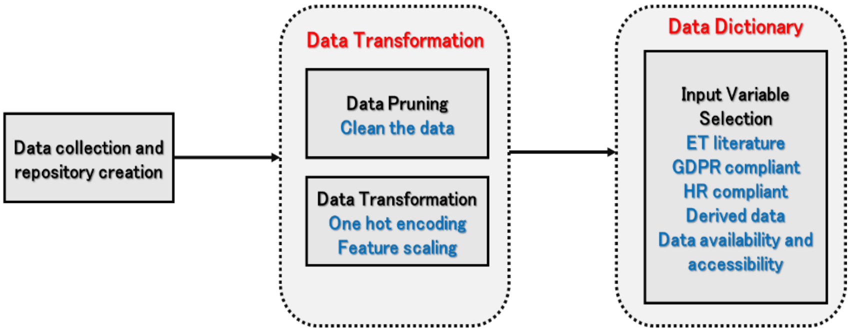


Figure 3: Data Preparation

Data Transformation prepares the dataset for the ML algorithm. The first transformation employed is single hot encoding converting all the categorical variables (text-based) into numerical data, consistent with the other columns. Next, scaling through normalisation is employed to transform the data into a specific range, ensuring a consistent numerical scale.

A *data dictionary* is created to examine and select each input variable considering the following principles: (1) can predict the outcome drawing upon the literature ET; (2) conforms to the data subject rights and data protection principles (Addis and Kutar, 2020); (3) readily available and accessible through HR information management systems; (4) is not discriminatory, i.e. conforms to the general HR employee regulations (Armstrong and Taylor, 2020); (5) replacing a cluster of derived input variables with a single variable, reducing data redundancy and duplication. A variable is deemed suitable as ‘input’ if all these principles are satisfied (shown in Appendix-2).

Machine learning phase (Figure 4)

After selecting the twenty-three input variables (see Appendix 2), we did not find any dependencies between these variables in the collinearity matrix, therefore, treated them as predictors for the ML algorithm. A *split-sample approach* (Dobbin and Simon, 2011) is used to divide the dataset into a training set (algorithm learns from this set), validation set (to evaluate the algorithm and fine-tune the hyperparameters, and finally select the most appropriate classification model based on accuracy; the algorithm does not learn from this data), and test set (provide an unbiased evaluation of the trained classifier algorithm making the predictions). The aim of the split is to ensure that the test dataset sample used to examine the performance of the classifier (ML algorithm) is independent of the training dataset to avoid *data leaks* (defined as unintentional leakage of signal into the validation and test sets). This will introduce bias and lead to the classifier overfitting (i.e., produces high accuracy), and does not reflect the true performance of the classifier (Kaufman et al., 2012). The proportion of split is 70% for training, 15% for validation and 15% for test. This is derived from the existing literature (recommended - two-thirds of the dataset should be used for training, where total number of cases < 5000) to train ML classifier algorithms (Bzdok et al., 2017; Kaufman et al., 2012).

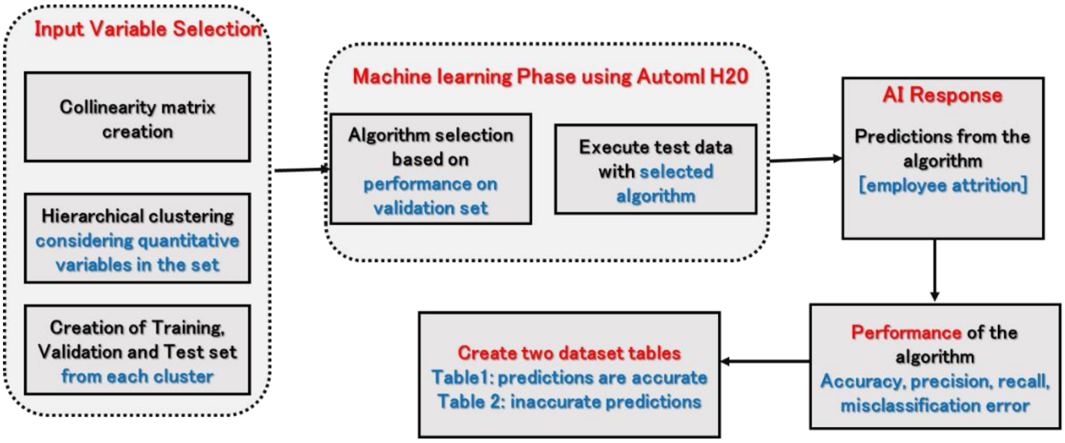

Input Variable Selection
Collinearity matrix creation
Hierarchical clustering considering quantitative variables in the set
Creation of Training, Validation and Test set from each cluster
Machine learning Phase using Automl H20
Algorithm selection based on performance on validation set
Execute test data with selected algorithm
AI Response
Predictions from the algorithm [employee attrition]
Performance of the algorithm
Accuracy, precision, recall, misclassification error
Create two dataset tables
Table1: predictions are accurate
Table 2: inaccurate predictions

Figure 4: Machine learning phase

Bias in the training sets due to *data-shift* (training data having different characteristics compared to the test data) may either lead to overfitting a model, i.e., the ML algorithm has high accuracy for the training set, while lower accuracy in the test sets, or underfitting, i.e., has lower accuracy for both the training and test datasets (Blitzer et al., 2007). We reduce bias in the training dataset through the following: (1) *hierarchical clustering* (Murtagh and Contereras, 2012) is employed on the whole dataset to group the cases together sharing similar characteristics considering the input variables in the whole dataset; we employ the split-sample approach on each cluster to ensure that the training dataset is representative of the whole dataset. This has two advantages (Bzdok et al., 2017): (1) the ML algorithm will potentially learn from all the characteristics and patterns; (2) the learning will enhance the performance of the algorithm over the test dataset, with one exception, i.e., it is unlikely to give good results for a new case not sharing similar characteristics to the training clusters. The solution is to use a semi-supervised ML algorithm (Van Engelen, 2020).

Next, *machine learning software package* (H20) is used to develop the classifier for predicting employee attrition. This software package comprises of commonly used ML algorithms to build classifiers. We found no statistically significant differences between the algorithmic performances. However, considering the overarching aim of this study, we have selected the *deep learning algorithm* because: (1) they are considered most opaque due to the presence of hidden layers between the input and output layers; (2) outputs produced by deep learning is extremely difficult to explain; (3) these algorithms are popular in real-life applications due to their ability to handle irrelevant features (i.e., separating signal from noise).

Model Explanation and Interpretation using LIME (Figure 5)

The desired characteristics of LIME explaining the output of any ML model are outlined below (Ribiero et al., 2016):

- *Interpretability:* it provides a qualitative and intuitive understanding of the relation between the input variables and the model's output response, which is easy to understand, without having any technical expertise of ML techniques.
- *Local fidelity:* it provides a complete and accurate description of the model by demonstrating the impact of input features in the vicinity of the instance being predicted, i.e. for each row of the data (each employee/case). At the same time, the model is able to provide a global interpretation, i.e. performance of the model for the whole dataset. This characteristic aids in understanding the behaviour of the model in various instances, which is necessary for managers to trust the model and examine its performance, not just statistically, but based on their social intelligence and contextual understanding.
- *Model-agnostic:* it treats every model as a black box and can explain any model. However, LIME has the ability to provide a high level overview of the model employing non-linear and complex classifiers, such as the one used in this study.

Therefore, LIME will help the managers to decide when to trust or not to trust the predictions made by the model. The execution of the LIME software package comprises of two phases (Ribiero et al., 2016): (1) *Explainer phase,* where the ML classifier is examined to understand the relationships between the input and output variables for each case. The key objective is to divide the test dataset into two classes based on the cases correctly and incorrectly predicted by the algorithm and then execute the explainer; (2) *Explanations phase* draws from the explainer response and presents them using tables and visual plots to help understand the rationale for the predictions in each case (key input variables contributing to the predictions) and why the model has underperformed (by visually looking at the explanations for cases predicted incorrectly).

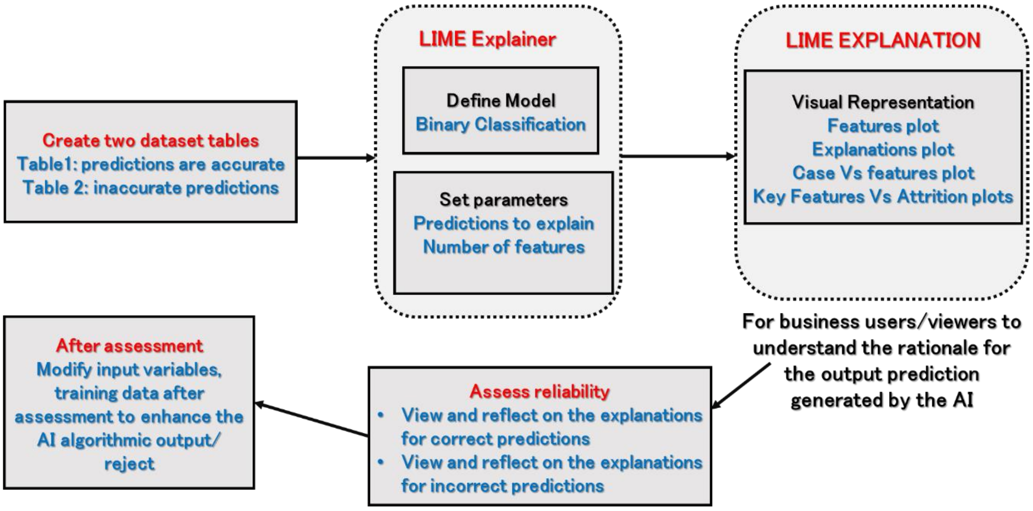
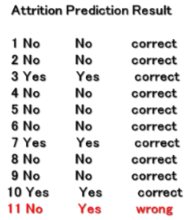

Create two dataset tables
Table1: predictions are accurate
Table 2: inaccurate predictions
LIME Explainer
Define Model
Binary Classification
Set parameters
Predictions to explain
Number of features
LIME EXPLANATION
Visual Representation
Features plot
Explanations plot
Case Vs features plot
Key Features Vs Attrition plots
For business users/viewers to understand the rationale for the output prediction generated by the AI
Assess reliability
• View and reflect on the explanations for correct predictions
• View and reflect on the explanations for incorrect predictions
After assessment
Modify input variables, training data after assessment to enhance the AI algorithmic output/ reject

Figure 5: Explanation phase

Sense-making Visualisation and Results

Sense-making of data visualization includes data transformation, representation, and interaction, which is ultimately about harnessing human visual perception capabilities to help identify trends, patterns, and outliers (Huhtamäki et al, 2015). Liu et al (2017) have suggested that the process of understanding, diagnosing, and refining machine learning models using visualization, is very important for users to efficiently solve ML problems. Visualisation can aid in understanding the ML models and output (Paiva et al, 2015) to diagnose model performance (Amershi et al, 2015) and refine the model (Liu et al, 2014). In the case of the predicting ET, visualisations can support the evidence from the ML algorithms as to why certain individuals may leave their employment, providing an overview of the quality of the base data and the strength for categorisations put forward. The prediction made by ML is shown for twenty employees in Figure 6. This output does not provide sufficient information about how and why these predictions were made. This lack of transparency will affect the trust of the managers on the ML predictions. The importance of the input variables at a local level (i.e., for each case/employee prediction) is unknown, which makes it difficult to devise retention strategies that will cater to the needs of the individuals. This brings into question, the importance and significance of designing transparency into the opaque machine learning algorithms, which will first aid the managers to understand the rationale for the algorithmic predictions, and then use their intuitive thinking and tacit domain experience to either accept or reject the prediction (Fountaine et al, 2019; Morse, 2020).


```
Attrition Prediction Result

 1 No      No      correct
 2 No      No      correct
 3 Yes     Yes     correct
 4 No      No      correct
 5 No      No      correct
 6 No      No      correct
 7 Yes     Yes     correct
 8 No      No      correct
 9 No      No      correct
10 Yes     Yes     correct
11 No      Yes     wrong
```

Figure 6: Output of the AI algorithm

The visual representations generated by the LIME explanation phase showing the key input variables that were used by the ML algorithm to generate the prediction for one example case is shown in Figure 7. Therefore, the manager now has a better and deeper understanding of the process followed to make the prediction. The key variables help to understand the reasons for ET at the local level (i.e., individual), and at the global level as well (i.e., for the whole sample used in the study).

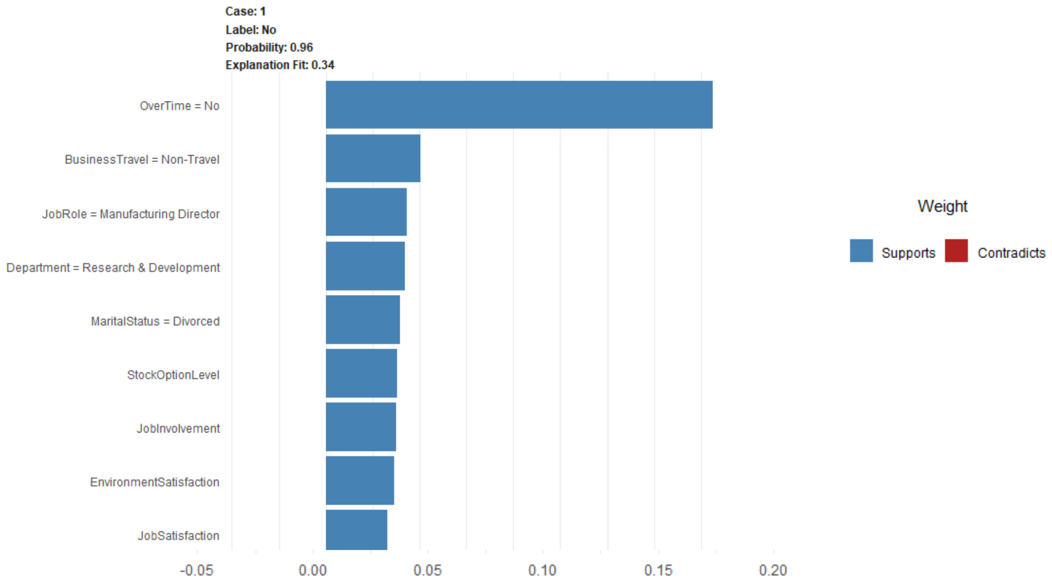


Figure 7: Visual representation of a case explained by LIME

Figure 8 shows the feature importance plot from LIME, visualizing the correct attrition predictions made for two employees. The top five input variables for each case will help the managers to understand: (1) how feature importance varies across each employee, i.e., the reason for turnover considering the individual information on each employee; (2) compare these features for a range of employees, through a filtering mechanism, to understand the accuracy and relevance of the model. The blue bars (Figure 9) mean that the features support the model conclusion, and the red bars contradict.

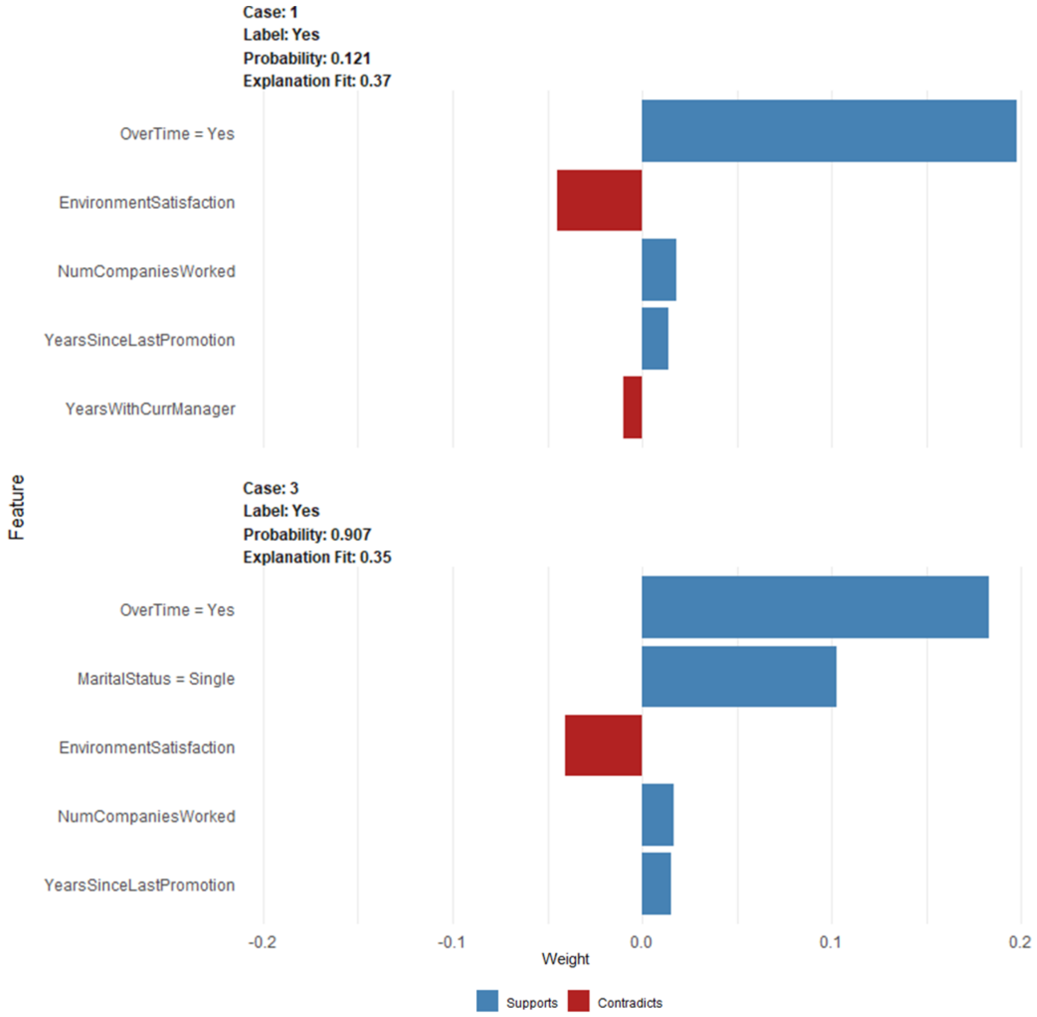


Figure 8: Comparing output explanation for two employees [correct predictions]

Figure 8 shows that the primary reason for predicting staff turnover as yes in case 1 is attributed to overtime by the algorithm, although there are other features such as environment satisfaction, years with manager, that contradict the prediction. From a managers' perspective, they can further drill-down and explore the cases, where staff turnover is attributed to overtime, and showing similar contradictions, to understand if this can be attributed to a specific job role and retention strategies to manage and retain talent in that specific role. Thus, the explanation will aid managers to uncover role and department specific retention issues, which may not have been possible without the AI model providing explanations on the process to decide reach a decision at a local level (i.e., for each individual employee). Figure 9 also shows that overtime, years since last promotion (development within the career) and number of companies worked (frequency of changing companies) are weighed high by the algorithm to predict the likelihood of attrition, which is also supported by the ET literature reviewed earlier (Lee, 2018; Farrington, 2008; Eisenberger et al., 2002).

However, from a manager's perspective to understand the reliability of the model, it is essential to understand the inaccurate predictions generated by the algorithm. Figure 9 shows the feature importance plot visualizing the case of two employees from the test dataset, where the model has incorrectly predicted the outcome variable attrition (as YES), although the label is NO, meaning that the staff did not leave the job. The algorithm predicts staff attrition as Yes, based on overtime, number of companies worked and the years since last promotion (for cases 1 and 3 in Figure 10), which is sensible and well-aligned to the literature and previous predictions (Timming, 2012; Farrington, 2008; Hackman and Oldham 1975). It seems that overtime is highly weighted by the model (a trend found across the dataset contributing to the attrition), which makes the final prediction inaccurate. Using this information, a HR manager can accurately understand the process followed by the algorithm and assess its reliability. They can also engage with the AI decision-support system to eliminate overtime as an input variable and then examine the reliability and accuracy of the model based on other input variables.

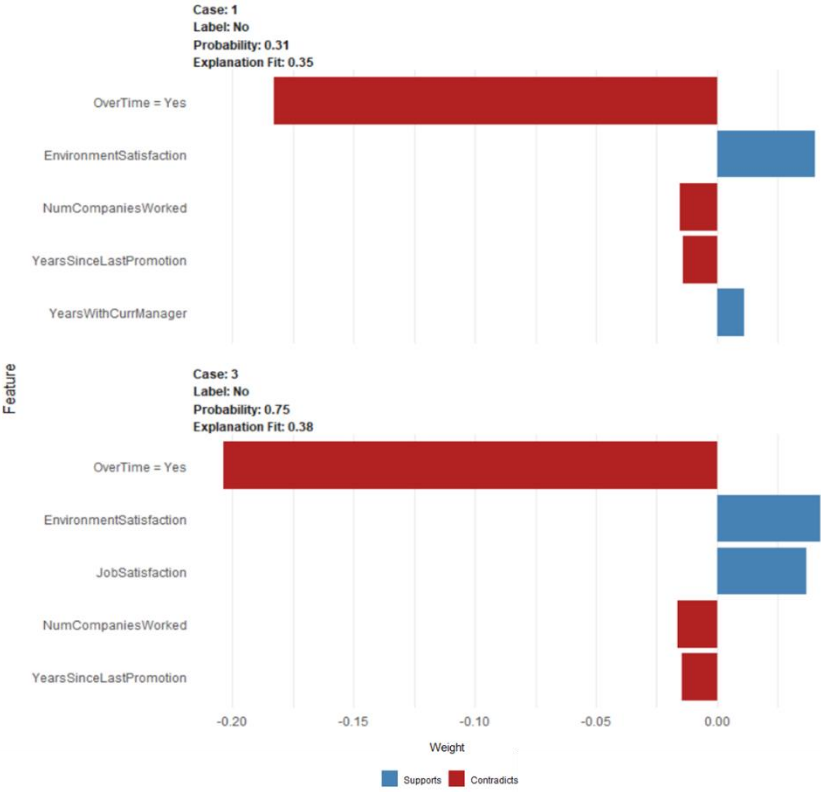


Figure 9: Comparing output explanation for two employees [incorrect predictions]

Managers can use their tacit experience and social intelligence (based on intuitive thinking), to determine the accuracy of the model for individual cases (Keding, 2020). Therefore, machine learning provides analytical intelligence to make strategies based on data-driven decisions, and the interpretability at the local level enhances the trust in the predictions made by the model (Ransbotham et al., 2020; Caputo et al., 2019; Jarrahi, 2018). Similarly, in the figure, we find another case, where a few parameters contradict the output, and these cases will require closer analysis from the managers, i.e., they are given an accurate reflection of how the model has made the predictions, and whether these predictions hold true in the given context. For example, the literature may report that working overtime will lead to high staff turnover (Thatcher et al., 2002; Cornell and Shapiro, 1987), however, this cannot be generalised for all job roles from the AI prediction (Alsheiabni, 2020; Glikson and Wooley, 2020).

The performance of the model is evaluated using Precision and Recall (Davis and Goadrich, 2006). Precision refers to the percentage of results predicted by the model, which are relevant, i.e., how many times the model predicts turnover as YES and it is YES, based on the training data set. On the other hand, recall is the percentage of total relevant results correctly predicted by the model, i.e., when the model predicts turnover as YES and how often it has correctly predicted it. The precision of the model is 62%, i.e., out of 100, 62 times the model has correctly predicted the response [YES]. The recall is 79%, i.e., if an organisation loses 100 employees, they can now target 79 of them (high probability of leaving). Therefore, from the viewpoint of a HR manager, they will be keen on recall value to focus on employees who are highly likely to leave the organisation, based on the prediction made by the model.

Discussion

Artificial intelligence provides analytical capabilities to enable data-driven decision-making for HR managers building on their social and strategic intelligence considering their interaction with socio-economic systems and tacit experience (Morse, 2020; Brock and Wangenheim, 2019). The three challenges that plague decision-making in organizations are: uncertainty, complexity, and equivocality (Choo, 1991; Simon, 1972). However, data-driven decision-making also requires creative vision, understanding of the convoluted social and political dynamics, and social mechanisms such as persuasion and negotiation to address uncertainty and equivocality. AI in its current state is unlikely to mimic human problem solving in these areas. Additionally, sense-making (Weick, 1995) and sense-giving are of utmost importance in the organisation from managers' perspectives. However, opaque AI systems and lack of AI literacy among the managers, makes it difficult to understand the rationale behind the automated recommendations and decisions made by AI systems (Choudhury, 2020).

While existing research in HRM has highlighted and indicated that collaborative intelligence stemming the combination of AI and human intelligence can enhance the quality of strategic decision-making (Jaiswal et al., 2021; Malik et al., 2019), our research demonstrates how AI transparency can help to drive organizationally valued outcomes (strategically managing ET). The findings from the preceding section provides evidence and valuable insights into how human intelligence (HI) and AI will offer distinct complementary qualities needed for effective decision-making to understand employee turnover and develop employee retention strategies. These insights pave the way for a partnership between HI and AI leveraging their unique capabilities (Iansiti and Lakhani, 2020). Data-driven decision-making can be best handled by using a blend of both analytical and intuitive approaches (Hung, 2003; Martin, 2009). Humans and AI can collaborate to deal with different aspects of decision-making, actively enhancing each other's complementary strengths: intuitive, empathetic, and interpersonal skills of the former, and analytical capabilities of the latter (Jarrahi, 2018). For example, AI can boost analytic abilities by processing vast amounts of data and providing the right information at the right time in the right place efficiently, using its superior computational abilities. HI can focus more on intuitive, empathetic, and interpersonal skills to deal with uncertainty and equivocality, and connect with people (negotiation and persuasion) by understanding the social fabrics of the organisation. Therefore, managers must restructure jobs and redesign tasks and roles, so that humans can spend more time on strategy development, empathetic and people-oriented tasks, leveraging on the analytical capabilities of AI systems.

Until now, most organisations have relied on HI to make decisions, which has indeed provided fruitful results and visions, over the past three decades (Davenport and Ronanki, 2018). In this context, every decision can be considered as a combination of analysis, deduction, and intuition (Burk and Miller, 1999). In data-driven decision-making, the insights gained from the data by employing analytics will drive deduction and intuition (Tambe et al., 2019; Caputo et al., 2019; Wilson et al., 2017). While AI-based machine learning models have become increasingly popular and important for data-driven decision-making in many areas of management, however, the process employed by the models to make predictions (classification and/or clustering), are often difficult for humans to understand and decipher. Many organisations fail to experience benefits from AI despite investing time, effort and resources, which is often attributed to limited digital skills, complex cognitive and information processing skills (Makarius et al., 2020). For instance, employees will require cognitive skills to interpret the AI output responses in a meaningful way (often referred to as sensemaking), which requires them to understand how these responses are generated by the AI system (Jarrahi, 2018). In this context, the interpretability of AI output responses resulting from embedding transparency will help the employees (and/or managers) to better understand the relevance of these outputs. Such understanding stemming from AI transparency will be instrumental for enhancing the trust in AI systems, enabling humans to take decisions in an efficient and more precise manner.

Given the proliferation of AI-based solutions in HRM business processes and practices, our research contributes a framework to embed transparency in AI-based solutions, which can increase the trust of the employees and managers, in these systems. Furthermore, AI transparency can help to mitigate issues related to fairness, discrimination, and trust, which have become increasingly important in managing human resource and subsequently making recruitment decisions as well as retention strategies.

Theoretical Implications

We contribute to the theory by adopting a RBV lens at the core of transparent AI-based ML models for ET in HR. This perspective is meant to address a fundamental challenge of using AI technology that can be hard to understand and trust, failing to become a key strategic resource for organisations. The adoption of RBV emphasises that technology becomes a strategic resource for sustainable competitive advantage through its use (Wernerfelt, 1984). Regardless of how good the decision support tool/model is, its purpose will fail without managerial trust and adoption. As a result, transparency in AI-based ML models is an enabler to confidently realise their value, overcoming trust issues related to AI outputs in the managerial decision-making domain, which is associated with the opaqueness of AI algorithms, leading to its limited use (Andriopoulos & Lewis, 2009; Raisch & Krakowski, 2021). The RBV perspective in this paper shifts the attention away from the adoption of new age technologies and tools (AI systems) towards trustworthiness in AI technology (through transparency and explainability), as a strategic intangible resource to achieve sustainable business competitiveness.

AI transparency will enhance the ability of HR managers within the organisations to understand, interpret and explain the automated outputs (i.e., ET predictions), which will equip them to create effective retention strategies. It will also increase managers' trust in the AI output responses, which will result in new knowledge creation to strategize both process and resource efficiency (Tambe et al., 2019). This will enhance the organisations' ability to develop capabilities to positively impact the productivity of employees and business (Makarius et al., 2020). Furthermore, flaws in the AI models can be easily identified, interpreted, and systematically mitigated, which will aid in improving the accuracy of the output responses (Satell and Sutton, 2019). This will lead to superior organizationally valued outcomes (correctly predicting staff turnover at a more granular level), and therefore, maximize the value created and captured by the organisations. Finally, AI transparency will also increase fairness and accountability, as managers can clearly explain the strategies, practices and policies derived from AI output responses to the workforce. This will enhance the confidence, trust and clarity, among the employees, which is likely to enhance employee engagement, psychological outcomes and commitment, resulting in low turnover (Brougham & Haar, 2018). Therefore, AI transparency will provide analytical agility (Barro and Davenport, 2019) and sustainable competitive advantage for organisational data-driven decision-making by strategically reducing employee turnover and enhancing workforce productivity, which will also positively impact business productivity.

Practical Implications

We further build on the tenets of RBV to outline the core competencies required within the HR department for the development and implementation of transparent AI systems, which will facilitate developing capability in the organisations through talent acquisition, management, and retention. The development team will require an array of skills and complementary viewpoints such as data scientists, AI expertise, visual analytics experts and HR domain expertise having a clear understanding of the business problem and motivation to employ AI and its benefits (Alsheiabni et al., 2020). The implementation team will require end-users (HR professionals who will be adopting the solution), and therefore in addition to domain expertise, they will require an understanding of the relevance of the data used in the analysis, and the accuracy of the recommendations to assess the effectiveness of algorithms' training process (Keding, 2020). Though demand for data scientists, machine learning experts, and robotic engineers are clearly growing, the importance of creativity, leadership, emotional intelligence, domain expertise and tacit experience are the key skills and competencies to drive AI transparency and subsequently evolve ML algorithms (Correani et al., 2020; Jarrahi, 2018). The human-AI collaboration resulting from AI transparency outlined in this paper identifies areas where AI can augment rather than replace humans in decision-making, and complementary intelligence can evolve AI models through explainability. We propose several recommendations for organisations.

- Organisations will need to first assess the decision-making tasks, next decide the core skills and competencies required to complete these tasks, and then make strategic decisions segregating the tasks between humans and AI. Such strategies will lead to the creation of new tasks, modifying the existing job descriptions, creating new roles, and modifying the organisation structure, rather than eliminating humans from the process (Gulliford and Dixon, 2019).
- Investments to train managers by providing AI literacy and skills will help companies to reap the benefits of human-AI symbiosis resulting from transparent AI. This will aid in consolidating and making the best use of human talents – creativity, communication, empathy, negotiation, intuition, persuasion and negotiation, which are necessary from the growth of the organisation and are currently the limitations of AI (Davenport & Bean, 2017).
- Organisations will need to appreciate and understand the capabilities of human intelligence in decision-making, to strike the right balance between investing in intelligent technologies and maintaining existing businesses (Davenport, 2016). Additionally, new procedures are required to govern AI, which will ensure the automated decisions conform to regulatory requirements and are ethical. This will lead to redefining and rethinking the decision-making process in terms of accountability, rewards, risks, investment, and its long-term sustainability (Kiron and Schrage, 2019).

Application of the Framework in Practice

In terms of employing the framework, a specific example case is General Practitioners (GPs). GPs are highly skilled professionals, who make complex decisions about patient care, require a high degree of knowledge and learning and have a skill level that requires accreditation. In 2014 there were around 37,000 full-time equivalent (FTE) GPs in England, working in around 7,875 practices (Kings Fund, 2016), with an increasing demand for their services (an increase of 6,610 to 7,171 patients per practice between 2010 and 2014) – (NHS, 2015). It is said that primary care is in crisis (Baird et al, 2016), due to increasing patient demand and a reduction in GPs, due to GPs leaving the profession early, retiring and there not being enough new recruits being trained. The consequence is substantial staff shortages, increased waiting times, and significant detrimental implications for patient care. Understanding and alerting management to impending attrition issues is an important factor in addressing staffing shortfalls and quality healthcare provision. The issue of understanding why GPs are leaving the profession is of pressing importance when the healthcare workforce is increasingly under pressure (Siddique, 2019). The ability to understand larger datasets using AI and machine learning will add to the evidence of attrition in GPs, and to understand what drives those attrition rates.

The existing research in HRM has discussed several important applications of AI such as recruitment and selection of applicants, video interviews, employee performance appraisal, talent prediction and coaching to reduce costs and enhance efficiency (Malik et al., 2020). Our proposed AI transparency framework can be applied in each of these applications to explain the AI output responses, which may help to reduce bias algorithms, models, and data inputs. This will facilitate developing AI systems for more efficient decision-making in the context of HR business processes and operations (Choudhury et al., 2020). Therefore, embedding transparency within the AI systems will aid HR decision-makers (or employees) to embrace, interact with and trust AI systems, which is critical for successfully adopting AI systems and experiencing the anticipated benefits (Makarius et al., 2020).

Conclusion

The globalization of multinational business enterprises and their operations has resulted in strategic management and retention of human resources as a critical factor contributing to overall organisational performance and productivity. Despite the interests and claims regarding the benefits offered by AI systems in HRM processes, developing a hybrid workforce, and re-conceptualising workplace culture, research on how AI transparency can be achieved is scant (Budhwar & Malik, 2019, 2020). In this context, to adopt AI in managerial decision-making, it is necessary to establish the foundations of trust in these systems, which has emerged as a focal point in HR and general management research. This article contributes to the IHRM literature by offering an implementation framework that demonstrates the use of LIME for explaining the employee turnover predictions made by AI-based ML models, thereby enhancing AI transparency and explainability. These explanations will enhance the reliability and trustworthiness of AI-based models for HR managers and employees, which will facilitate the effectiveness and efficiency of data-driven strategic HR decision-making to create sustainable value in business organisations.

The framework proposed in this article can be empirically tested with HR managers and in other business sectors to examine the drivers and barriers to designing transparent AI systems. Future research should examine, when and how organisations can switch from opaque to transparent AI, to better understand the automated decisions made by AI, conforming with the policies set by the regulators. AI transparency frameworks will require collaborative intelligence, therefore future research should also examine and re-structure the roles and responsibilities of humans and technology in data-driven decision-making. Finally, researchers should develop frameworks to successfully integrate transparent and interpretable AI within organisational processes, considering the joint optimization of an organization’s human and technology capability within a given context.

References

Addis, C. and Kutar, M., 2020. General Data Protection Regulation (GDPR), Artificial Intelligence (AI) and UK Organisations: A year of implementation of GDPR.

Agarrwal et al., 2020. How to Win with Machine Learning. Harvard Business Review

Alalie, H.M., Harada, Y. and Noor, I.M., 2018. A Resource-Based View: How Information Technology Creates Sustainable Competitive Advantage to Improve Organizations. *Journal of Advance Management Research*, *6*(12), pp.1-5.

Allen, D. G. (2008) Retaining talent: A guide to analyzing and managing employee turnover. SHRM Foundations.

Alsheiabni, S., Messom, C., Cheung, Y. and Alhosni, M., 2020. Winning AI Strategy: Six-Steps to Create Value from Artificial Intelligence.

Amabile, T.M., 2020, Creativity, Artificial Intelligence, and a World of Surprises. Academy of Management Discoveries, 6(3).

Amershi, S., Chickering, M., Drucker, S.M., Lee, B., Simard, P. and Suh, J., 2015, April. Modeltracker: Redesigning performance analysis tools for machine learning. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 337-346).

Ananny, M. and Crawford, K., 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), pp.973-989.

Andriopoulos, C. and Lewis, M.W., 2009. Exploitation-exploration tensions and organizational ambidexterity: Managing paradoxes of innovation. *Organization science*, *20*(4), pp.696-717.

Baird, B., Charles, A., Honeyman, M., Maguire, D. and Das, P., 2016. Understanding pressures in general practice. London: King's Fund.

Barney, J.B., 2001. Resource-based theories of competitive advantage: A ten-year retrospective on the resource-based view. Journal of management, 27(6), pp.643-650.

Barro, S. and Davenport, T.H., 2019. People and machines: Partners in innovation. MIT Sloan Management Review, 60(4), pp.22-28.

BBC, 2020. Apple's 'sexist' credit card investigated by US regulator. Accessed on 20 Sep 2020, Available at https://www.bbc.com/news/business-50365609

Bieda, L., 2020. How Organizations Can Build Analytics Agility. MIT Sloan Management Review

Blitzer, J., Dredze, M. and Pereira, F., 2007, June. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 440-447).

Boxall, P., 1996. The strategic HRM debate and the resource-based view of the firm. Human resource management journal, 6(3), pp.59-75.

Brock, J.K.U. and Von Wangenheim, F., 2019. Demystifying AI: What digital transformation leaders can teach you about realistic artificial intelligence. California Management Review, 61(4).

Bromiley, P. and Rau, D., 2016. Operations management and the resource-based view: Another view. Journal of Operations Management, 41, pp.95-106.

Brougham, D. and Haar, J., 2018. Smart technology, artificial intelligence, robotics, and algorithms (STARA): Employees' perceptions of our future workplace. *Journal of Management & Organization*, *24*(2), pp.239-257.

Budhwar, P. and Malik, A., 2020. Special Issue: Leveraging Artificial and Human Intelligence through Human Resource Management. Call for paper), Human Resource Management Review.

Budhwar, P.S. and Malik, A., 2019. Call for papers: Artificial intelligence challenges and opportunities for international HRM. The International Journal of Human Resource Management.

Bzdok, D., Krzywinski, M. and Altman, N., 2017. Machine learning: a primer.

Caputo, F., Cillo, V., Candelo, E. and Liu, Y., 2019. Innovating through digital revolution: The role of soft skills and big data in increasing firm performance. Management Decision.

Caves, R.E., 1980. Industrial organization, corporate strategy and structure. In *Readings in accounting for management control* (pp. 335-370). Springer, Boston, MA.

Chatterjee, J., 2020. AI beyond accuracy: Transparency and Scalability. Accessed on 01 Dec 2020. Available at: https://towardsdatascience.com/ai-beyond-accuracy-transparency-and-scalability-d44b9f70f7d8

Cheng, M.M. and Hackett, R.D., 2019. A critical review of algorithms in HRM: Definition, theory, and practice. Human Resource Management Review, p.100698.

Choo, C.W., 1991. Towards an information model of organizations. The Canadian Journal of Information Science, 16(3), pp.32-62.

Choudhury, P., Starr, E. and Agarwal, R., 2020. Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, *41*(8), pp.1381-1411.

Chowdhury et al., 2020. Putting Responsible AI Into Practice. MIT Sloan Management Review

Ciechanowski, L., Jemielniak, D. and Gloor, P.A., 2020. TUTORIAL: AI research without coding: The art of fighting without fighting: Data science for qualitative researchers. Journal of Business Research, 117, pp.322-330.

Collins, C.J., 2020. Expanding the resource-based view model of strategic human resource management. The International Journal of Human Resource Management, pp.1-28.

Correani, A., De Massis, A., Frattini, F., Petruzzelli, A.M. and Natalicchio, A., 2020. Implementing a digital strategy: Learning from the experience of three digital transformation projects. California Management Review, 62(4), pp.37-56.

Cornell, B. and Shapiro, A.C., 1987. Corporate stakeholders and corporate finance. Financial management, pp.5-14.

Cowgill, B. and Tucker, C.E., 2020. Algorithmic fairness and economics. *The Journal of Economic Perspectives*.

Crain, M (2018) The limits of transparency: data brokers and commodification. New Media & Society 20(1): 88–104.

Davenport, T.H. and Ronanki, R., 2018. Artificial intelligence for the real world. Harvard business review, 96(1), pp.108-116.

Davenport, T. H., & Bean, R., 2017. How P&G and American Express are approaching AI. Harvard Business Review. Accessed on 20 Sep 2020. Available at: https://tinyurl.com/sf9uvuv4

Davenport, T.H., 2016. Rise of the strategy machines. MIT Sloan Management Review, 58(1), p.29.

Davis, J. and Goadrich, M., 2006, June. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

Degbey, W.Y., Rodgers, P., Kromah, M.D. and Weber, Y., 2020. The impact of psychological ownership on employee retention in mergers and acquisitions. *Human Resource Management Review*, p.100745.

Dobbin, K.K. and Simon, R.M., 2011. Optimally splitting cases for training and testing high dimensional classifiers. BMC medical genomics, 4(1), p.31.

Dougherty, C (2015) Google photos mistakenly labels black people "gorillas." The New York Times, Accessed on 1 July 2020. Available at: https://tinyurl.com/yk7p7wd6

Edgar, F., Geare, A. and Zhang, J.A., 2017. A comprehensive concomitant analysis of service employees' well-being and performance. *Personnel Review*.

Eisenberger, R., Stinglhamber, F., Vandenberghe, C., Sucharski, I.L. and Rhoades, L., 2002. Perceived supervisor support: contributions to perceived organizational support and employee retention. *Journal of applied psychology*, *87*(3), p.565.

Farrington, R.E., 2008. Mission: reduce employee turnover; game plan: employee education. *Revenue-cycle strategist*, *5*(12), p.6.

Fountaine, T., McCarthy, B. and Saleh, T., 2019. Building the AI-powered organization. *Harvard Business Review*, *97*(4), pp.62-73.

Glikson, E. and Woolley, A.W., 2020. Human trust in Artificial Intelligence: Review of empirical research. Academy of Management Annals, (ja).

Gulliford, F. and Dixon, A.P., 2019. AI: the HR revolution. Strategic HR Review.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S.F., Childe, S.J., Hazen, B. and Akter, S., 2017. Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, 70, pp.308-317.

Gunning, D., 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*(2).

Gusnadi, Y. and Hermawan, A., 2020. Designing Employee Performance Monitoring Dashboard Using Key Performance Indicator (KPI). *bit-Tech*, *2*(2), pp.19-26.

Haenlein, M. and Kaplan, A., 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California management review, 61(4), pp.5-14. ange, 162, p.120392.

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S. and Broderick, T., 2020. Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), pp.E14-E16.

Huhtamäki, J., Russell, M.G., Rubens, N. and Still, K., 2015. Ostinato: The exploration-automation cycle of user-centric, process-automated data-driven visual network analytics. In *Transparency in social media* (pp. 197-222). Springer, Cham.

Hung, S.Y., 2003. Expert versus novice use of the executive support systems: an empirical study. Information & Management, 40(3), pp.177-189.

Huselid, M.A., 2018. The science and practice of workforce analytics: Introduction to the HRM special issue. Human Resource Management, 57(3), pp.679-684.

Hwang, S., Kim, J., Park, E. and Kwon, S.J., 2020. Who will be your next customer: A machine learning approach to customer return visits in airline services. Journal of Business Research, 121, pp.121-126.

Iansiti, M. and Lakhani, K.R., 2020. Competing in the Age of AI. Harvard Business Review

Jaiswal, A., Arun, C.J. and Varma, A., 2021. Rebooting employees: upskilling for artificial intelligence in multinational corporations. The International Journal of Human Resource Management, pp.1-30.

Jarrahi, M.H., 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4), pp.577-586.

Jergensen, G.E., McGovern, A., Lagerquist, R. and Smith, T., 2020. Classifying convective storms using machine learning. Weather and Forecasting, 35(2), pp.537-559.

Johnson, R.D., Stone, D.L. and Lukaszewski, K.M., 2020. The benefits of eHRM and AI for talent acquisition. Journal of Tourism Futures.

Ju, B. and Li, J. (2019) 'Exploring the impact of training, job tenure, and education-job and skills-job matches on employee turnover intention', European Journal of Training and Development. University of Illinois at Urbana-Champaign, Champaign, IL, United States: Emerald Group Publishing Ltd., 43(3–4), pp. 214–231.

Kaggle. 2020. IBM data. Accessed on 5 March 2020 Available at: https://tinyurl.com/y32xdv5k

Kaplan, A. and Haenlein, M., 2020. Rulers of the world, unite! The challenges and opportunities of artificial intelligence. Business Horizons, 63(1), pp.37-50.

Kaufman, S., Rosset, S., Perlich, C. and Stitelman, O., 2012. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), pp.1-21.

Keding, C., 2020. Understanding the interplay of artificial intelligence and strategic management: four decades of research in review. Management Review Quarterly, pp.1-44.

Kiron, D., and Schrage, M., 2019. Strategy For and With AI. MIT Sloan Management Review

Kings Fund (2016). Understanding pressures in general practice. Accessed on 05 March 2020. Available at: https://tinyurl.com/284n2y8n

Krause, J., Perer, A. and Ng, K., 2016, May. Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5686-5697).

Lee, S. (2018) 'Employee Turnover and Organizational Performance in U.S. Federal Agencies', American Review of Public Administration. Indiana University BloomingtonIN, United States: SAGE Publications Inc., 48(6), pp. 522–534.

Lee, T.H. and Boynton, L.A., 2017. Conceptualizing transparency: Propositions for the integration of situational factors and stakeholders' perspectives. *Public Relations Inquiry*, *6*(3), pp.233-251.

Liu, C., Gong, S., Loy, C.C. and Lin, X., 2012, October. Person re-identification: What features are important?. In European Conference on Computer Vision (pp. 391-401). Springer, Berlin, Heidelberg.

Liu, S., Wang, X., Liu, M. and Zhu, J., 2017. Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics, 1(1), pp.48-56.

Liu, S., Wang, X., Chen, J., Zhu, J. and Guo, B., 2014, October. Topicpanorama: A full picture of relevant topics. In 2014 IEEE Conference on Visual Analytics Science and Technology (VAST) (pp. 183-192). IEEE.

Makarius, E.E., Mukherjee, D., Fox, J.D. and Fox, A.K., 2020. Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. Journal of Business Research, 120, pp.262-273.

Malik, A., Budhwar, P., Patel, C. and Srikanth, N.R., 2020. May the bots be with you! Delivering HR cost-effectiveness and individualised employee experiences in an MNE. The International Journal of Human Resource Management, pp.1-31.

Malik, A., Pereira, V. and Tarba, S., 2019. The role of HRM practices in product development: Contextual ambidexterity in a US MNC's subsidiary in India. The International Journal of Human Resource Management, 30(4), pp.536-564.

Malone, T.W., 2018. How human-computer 'Superminds' are redefining the future of work. MIT Sloan Management Review, 59(4), pp.34-41.

Martin, R.L., 2009. The design of business: Why design thinking is the next competitive advantage. Harvard Business Press.

Mikalef, P. and Gupta, M., 2021. Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Information & Management, 58(3), p.103434.

Morse G., 2020. Harnessing Artificial Intelligence. Harvard Business Review.

Murtagh, F. and Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp.86-97.

Neeley, R., 2017. Engagement Strategies to Reduce Registered Nurse Turnover in Hospitals.

NHS (2015). General and personal medical services, England – 2004-2014. Accessed on 24th January 2020. Accessible at: https://tinyurl.com/2csykbzk

O'Connell, M. and Kung, M.C., 2007. The Cost of Employee Turnover. *Industrial Management*, *49*(1).

O'neil, C., 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books

Paiva, J.G.S., Schwartz, W.R., Pedrini, H. and Minghim, R., 2014. An approach to supporting incremental visual data classification. IEEE transactions on visualization and computer graphics, 21(1), pp.4-17.

Pigni, F., Piccoli, G. and Watson, R., 2016. Digital data streams: Creating value from the real-time flow of big data. California Management Review, 58(3), pp.5-25.

Pillai, R. and Sivathanu, B., 2020. Adoption of artificial intelligence (AI) for talent acquisition in IT/ITeS organizations. Benchmarking: An International Journal.

Raisch, S. and Krakowski, S., 2021. Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, *46*(1), pp.192-210.

Rampersad, G., 2020. Robot will take your job: Innovation for an era of artificial intelligence. *Journal of Business Research*, *116*, pp.68-74.

Ransbotham, S., Khodabandeh, S., Fehling, R., LaFountain, B. and Kiron, D., 2019. Winning with AI. MIT Sloan Management Review, 61180.

Ransbotham et al., 2020. Expanding AI's impact with Organisational Learning,

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Satell, G. & Sutton, J. (2019). We need AI that is explainable, auditable and transparent. Harvard Business Review. Accessed on 10 January 2021. Available at: https://tinyurl.com/s5pyn7a

Shaw, J. D. et al. (2005) ‘Turnover, social capital losses, and performance’, Academy of management Journal. Academy of Management Briarcliff Manor, NY 10510, 48(4), pp. 594–606.

Shin, D. and Park, Y.J., 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior, 98,* pp.277-284.

Shrestha, Y.R., Krishna, V. and von Krogh, G., 2020. Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. Journal of Business Research, 123, pp.588-603.

Shrestha, Y.R., Ben-Menahem, S.M. and Von Krogh, G., 2019. Organizational decision-making structures in the age of artificial intelligence. California Management Review, 61(4), pp.66-83.

Siddique, H. (2019). NHS patients waiting over two weeks to see a GP, shows survey. Accessed on 05 March 2020. Available at: https://tinyurl.com/k7s69vs6

Accessible at https://www.theguardian.com/society/2019/aug/12/nhs-patients-waiting-over-two-weeks-to-see-a-gp-shows-survey Accessed on 05 March 2020

Silvernam, K., 2020. Why Your Board Needs a Plan for AI Oversight. MIT Sloan Management Review.

Stevens et al., 2020. How one Boston hospital built a COVID-19 forecasting system. Harvard Business Review. Accessed on. 01 Dec 2020.

Suen, H.Y., Chen, M.Y.C. and Lu, S.H., 2019. Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? Computers in Human Behavior, 98, pp.93-101.

Tambe, P., Cappelli, P. and Yakubovich, V., 2019. Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), pp.15-42.

Timming, A.R., 2012. Tracing the effects of employee involvement and participation on trust in managers: An analysis of covariance structures. The International Journal of Human Resource Management, 23(15), pp.3243-3257.

Van Engelen, J.E. and Hoos, H.H., 2020. A survey on semi-supervised learning. Machine Learning, 109(2), pp.373-440.

von Krogh, G., 2018. Artificial intelligence in organizations: new opportunities for phenomenon-based theorizing. Academy of Management Discoveries.

Wade, L (2010) HP software doesn’t see black people. Sociological Images. Accessed on 5 January 2021. Available at: https://tinyurl.com/cwccd9dc

Weick, K.E., 1995. Sensemaking in organizations (Vol. 3). Sage.

Wernerfelt, B., 1984. A resource-based view of the firm. Strategic management journal, 5(2), pp.171-180.

Winter, S.G., 2003. Understanding dynamic capabilities. Strategic management journal, 24(10), pp.991-995.

Zhang, Y., Song, K., Sun, Y., Tan, S. and Udell, M., 2019. why should you trust my explanation?” understanding uncertainty in lime explanations. arXiv preprint arXiv:1904.12991.

Appendix 1

Table 1 AI algorithms used for classification

Algorithms	**Description**	**Accuracy**
Logistic Regression	Using logistic function (sigmoid curve), the predictions are mapped to be between 0 and 1, represented as probability of occurrence in each class (where there are 'n' classes)	This can give high accuracy for small datasets but not suitable for use in big data sets having multiple dimensions.
Decision Tree	Employs a hierarchical tree structure, leaves represent the labels of the class, branches represent the features that lead to those class labels based on a given set of pre-defined rules.	The accuracy is similar to that of logistic regression, but it is not suitable for big data sets and analysing complex relationships between the input and outcome variables.
Support Vector Machine	Uses the concept of kernels (dot product between the input and output variables in the classification space). The aim is to maximize the distance between the closest members of distinct classes, through distance calculation in Euclidean space.	The accuracy is higher compared to most algorithms when used over small data sets, but lower compared to deep learning algorithms. This requires dimensionality reduction to analyse complex relationships between variables
Naïve Bayes	Based around the concept of conditional probability, where the model is a probability table for each case.	Accuracy is similar to logistic regression and decision trees; however, it is not suitable for use to analyse and predict from big data sets.
Random Forest	Employs building ensemble decision trees and merges them together to reach a decision point, i.e., prediction. Effectively produces feature importance plots and prevents overfitting through cross validation	Accuracy is higher compared to the above algorithms; however, the process is slower due to ensemble nature of the algorithm.
Gradient Boosting	Unlike random forest, this technique builds dependency trees i.e. uses the tress created during each iteration to build the next tree, therefore the results and rules are combined throughout the iterative process.	Accuracy is higher compared to decision trees and random forest, due to dependency between the iteration, however, the parameters are harder to tune compared to other algorithms.
Deep Learning	Based on neural network architecture by introducing hidden layers between the input and output response and automatically extracts features from the datasets.	Accuracy is very high for labelled datasets – supervised machine learning. Accuracy increases with the size of the dataset and is suitable for big data analysis

Appendix-2

Table 2 Data dictionary and input variable selection

Variables	Relevant to HR Literature	GDPR Compliant	Accessible	HR Compliant	Input Variable for prediction
BusinessTravel Frequency	Yes	Yes	Yes	Yes	Yes
Department	Yes	Yes	Yes	Yes	Yes
DistanceFromHome	Yes	Yes	Yes	Yes	Yes
Education	Yes	Yes	Yes	Yes	Yes
EducationField	Yes	Yes	Yes	Yes	Yes
EnvironmentSatisfaction	Yes	Yes	Yes	Yes	Yes
JobInvolvement	Yes	Yes	Yes	Yes	Yes
JobLevel	Yes	Yes	Yes	Yes	Yes
JobRole	Yes	Yes	Yes	Yes	Yes
JobSatisfaction	Yes	Yes	Yes	Yes	Yes
MaritalStatus	Yes	Yes	Yes	Yes	Yes
NumCompaniesWorked	Yes	Yes	Yes	Yes	Yes
OverTime	Yes	Yes	Yes	Yes	Yes
PercentSalaryHike	Yes	Yes	Yes	Yes	Yes
PerformanceRating	Yes	Yes	Yes	Yes	Yes
TotalWorkingYears	Yes	Yes	Yes	Yes	Yes
TrainingTimesLastYear	Yes	Yes	Yes	Yes	Yes
YearsAtCompany	Yes	Yes	Yes	Yes	Yes
YearsSinceLastPromotion	Yes	Yes	Yes	Yes	Yes
YearsInCurrentRole	Yes	Yes	Yes	Yes	Yes
YearsWithCurrManager	Yes	Yes	Yes	Yes	Yes
Daily Rate	Yes	Yes	Yes	Yes	Yes, select hourly rate, as other rates can be derived from this
HourlyRate	Yes	Yes	Yes	Yes	
MonthlyIncome	Yes	Yes	Yes	Yes	
MonthlyRate	Yes	Yes	Yes	Yes	
RelationshipSatisfaction	Yes	Yes	Not necessarily	Not necessarily	No
Gender	Yes	Yes	Yes	Yes (globally)	Yes (globally)
Age	No	Yes	Yes	Yes (globally)	Yes (globally)
EmployeeNumber	This is an ID pseudonym for each employee				No
Attrition	This is the outcome variable (prediction)				No

Data Availability Statement

The data that support the findings of this study are openly available under creative commons license and freely downloadable in [KAGGLE] at [https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset], referenced in this article as below.

Kaggle. 2020. IBM data. [ONLINE] Available at: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset. [Accessed 5 March 2020].