International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 1
DOI: https://doi.org/10.15157/IJITIS.2026.9.1.115-136
Received: 22.11.2025; Revised: 30.12.2025; Accepted: 12.01.2026

Research Article

Data-Driven Predictive Modelling of Employee Absenteeism Using Workflow Automation Platforms

Mohammed Alars[1]*, Abbas Albakry[2]

[1]Department of Computer Science, University of Information Technology and Communications, Baghdad, Iraq

[2]Department of Artificial Intelligence, University of Information Technology and Communications, Baghdad, Iraq

***mohamed.amnya@gmail.com**

Abstract

Employee absence is a critical factor affecting organizational productivity and employee well-being. This study presents a data-driven predictive framework for employee absenteeism using a newly collected enterprise dataset comprising 8,336 employees. Absenteeism is formulated as a binary classification task, distinguishing employees with more than 80 hours of annual absence from those with lower absence levels, based on demographic and occupational characteristics. The proposed approach applies gradient-boosted decision tree models, including LightGBM, XGBoost, and CatBoost, evaluated through a stratified train–test split at the employee level to approximate temporal separation between training and prediction. Feature engineering procedures are detailed, including categorical encoding and the construction of a commuting-related indicator. All models demonstrate strong predictive performance, achieving accuracy between 85% and 87%, precision ranging from 78% to 80%, recall between 76% and 79%, and AUC–ROC values of 0.92–0.93. Model interpretability is addressed using SHAP-based feature attribution, identifying age, gender, and occupational role and location as key predictors of absenteeism risk. Furthermore, a practical system architecture is outlined, integrating the predictive models within an automated workflow using the n8n orchestration platform for deployment in human resource information systems. This enables proactive identification of high-risk absenteeism cases and supports early intervention strategies with minimal human oversight. The study contributes by addressing data leakage concerns, improving feature transparency, and demonstrating a deployable and interpretable predictive system. Future research directions include multi-organizational validation, temporal modelling using sequential data, and evaluation of system-level effectiveness in real-world HR settings.

Keywords: Employee Absenteeism, Agent, LightGBM, XGBoost, CatBoost, Predictive Modelling, Human Resources, Machine Learning

INTRODUCTION

Employee absenteeism defined as unscheduled absence from work poses a significant challenge to organizations worldwide. Higher absenteeism can disrupt operations and

incur substantial costs, including reduced productivity, overtime, and decreased employee morale [1]. Apart from pressing financial issues, absence also creates concerns relating to health, engagement, and working conditions of associated staff members [2]. The combined effects emphasize its relevance as a subject for concern by both organizational management as well as health professionals.

The methods being used traditionally for absenteeism management at work are more reactive, dealing with absenteeism once it becomes a problem. Of late, there has also been an increased interest in using predictive analysis that would allow proactive management. Organizations are using historical data from HR as well as techniques like machine learning to identify which employees are at a higher risk of absenteeism and then take preventive steps before it leads to a problem. It has become possible due to predictive analysis, which has enabled early detection of employees who are at a risk of absenteeism [3].

Although many studies have been conducted on the prediction of absenteeism, previous research has faced challenges that this research study attempts to solve. Many previous research works were conducted using one public data set called "Absenteeism at Work" provided by UCI, which consists of only 740 instances [4, 5]. Since this data set was widely used, models trained on this data set have shown unfeasibly high accuracy (often higher than 95%) for certain reasons like leakage of data or misuse of data split [6, 7]. For instance, when records for the same person were split into both training and test data, some models might have tended to show high accuracy [8, 9].

In this work, we improve an agent-based predictive model on employee absenteeism. We use novel data, called *MFGEmployees*, comprising thousands of data points from an employee base in the manufacturing/retail sector. The predictive task is posed as a strictly defined binary classification task: to predict those workers that have high rates of "absenteeism" above a certain number in the allotted timeframe compared to those that have normal rates. Through deft data preprocessing to avoid any leaking information (e.g., variables noting rates of employee absenteeism), along with an assessment strategy that seeks to simulate actual utilization (temporal/employee split), we validate the models' predictive ability [10].

The strategy uses three gradient boosting algorithms, LightGBM, XGBoost, and CatBoost, which are chosen based on their ability to handle structured data classification effectively. Details are provided regarding feature engineering, preprocessing, and transformation necessary with each algorithm, such as encoding categorical variables such as department or employment role, transforming continuous variables, or using commuting distance, which is introduced with a new, simplified feature. The results are validated by comparing each model through standard performance calculations (Accuracy, Precision, Recall, or F1 Score, AUC) to determine predictive accuracy. To better interpret model results, SHAP (SHapley Additive exPlanations) value analysis is used to examine model output, which maps influential input variables explaining model predictions of high absenteeism. This information is critical to HR professionals wanting

to correctly act on model predictions to determine high absenteeism associations with tenure, age, or other workplace-related variables. Aware that a predictive model will only benefit the organization in which it is implemented, the next section of this paper discusses how the predictive model can be integrated with an automation environment using the n8n platform, which is an open-source automation tool preferred by many users. This section will present an agent-based implementation framework whereby an autonomous workflow will schedule the retrieval of updated HR information and the predictive model will automatically take the necessary steps (for instance, alerting the manager or organizing a check-up visit, among other interventions). This is how the "last mile problem" in analytics is overcome.

In this paper, there are three aspects where the contributions of this research lie: (1) providing a strong absenteeism prediction model using a large and modern dataset and better methodological practices, (2) gaining understanding of important absenteeism predictors through SHAP explainability, and (3) providing a guide on implementing the model within an automated agent-based HR solution. The findings show that with good dataset construction and testing, high predictive performance of over 85% to 90% accuracy and an AUC of 0.92 or better is possible without relying on black-box features and that this model can be effectively used to actively decrease absenteeism. The remainder of this paper is divided as follows: Section Related Work discusses recent studies within absenteeism prediction and management, Section Dataset Description introduces this new dataset and its target variable, Section Methodology describes this modelling process and experimentation, Section Results and Discussions presents model predictions and analyses, Section SHAP-based Interpretability investigates feature and explanation importance, Section Deployment Architecture elucidates its feasible use within an automation and agent-based solution, and finally, Section Conclusion and Future Work summarizes important findings, limitations, and future work directions.

RELATED WORK

The predictive model of absenteeism has also attracted intense research interests in the last decade, including traditional statistical models, as well as newer models in AI. The earlier research mainly focused on isolating demographic and occupational variables associated with absenteeism. Authors in [11] considered absenteeism in an Italian multi utility firm, where they identified department and staff variables as key drivers of absenteeism. Similarly, meta-analyses in organizational psychology research have associated organizational stress, as well as health concerns, with increased levels of absenteeism, thereby confirming that absenteeism is multi-faceted, involving individual, occupation, and environmental variables [5]. Table 1 depict the summary and comparison of related work.

Most specifically, a pilot study initiated by authors [12] used their particular dataset and attempted to predict whether an absence was of a prolonged type (>30 hours). They used both RF and GBM models to predict this classification problem and could reach an

accuracy of 84% and an AUC of 0.89 and concluded that their models stated that absence reasons (using Categorical ICD 10 Codes), Body Mass Index, and Workload could predict an absence of prolonged duration accurately. Most interestingly, they also indicated that other parameters, namely the time of absence and distance to work, had a non-trivial effect on absenteeism prediction, with approximately 11.8% and 9.8% importance of absence time and distance to work, then other times of an association and so forth [2].

Table 1. Summary and comparison of related work

Study	Methodology	Dataset Size	Domain	Key Features Used	Accuracy / AUC	Unique Contribution
[1]	Deep Learning (MLP)	~3,500 agents	Public Safety	Demographics, history	~91% / 0.89	Applied deep models to public sector
[2]	GBM, SVM, Ensemble	~1,200	Occupational Health	Distance, workload	88% / 0.89	Used AUC-PR and expert-verified labels
[3]	Clustering Trees, RF	~4,000	Retail	Attendance patterns	89% / NA	Short/long-term prediction with clusters
[4]	Neuro-Fuzzy	740	Courier	Leave type, tenure	~82% / NA	Early hybrid model on small data
[5]	Logistic Regression	Case Study	Multi-utility	HR logs, job category	NA	Management analysis
[6]	Logistic Regression, RF	~2,000	Education	Socioeconomic, tenure	~77%	Linked absenteeism to income
[7]	Feature selection + RF	~3,500 agents	Public Safety	Optimized subset of features	~93%	Ablation study and selection
[8]	Timesheet + Sliding Window	~2,500	IT/Retail	Timestamped logs	85–87%	Explored temporal effects
[9]	Stress-Productivity Theory	Theoretical	Cross-sector	Job strain, psychological factors	NA	Introduced job demand–control model
Our Study	LightGBM, CatBoost, XGBoost	8,336	Retail + Factory	SHAP, SameCity, Division	85–87% / 0.93	Scalable agent-based deployment via n8n

Though the initial modelling could be easily performed because of the presence of the UCI dataset, there have been concerns mentioned in the literature. The small scale and focused areas in the dataset have triggered overfitting in a number of studies as well as the saturation of knowledge there have been a plethora of papers since 2023 that have reused the same dataset with incremental improvements [4, 5]. Moreover, there have been a

number of features in the existing dataset (such as the cumulative absence time) that can result in information leakage if handled wrongly [7, 9]. As a response to these issues in the literature, our task will be to provide a new dataset with more than 8,000 employees along with ensuring that there are no features leading to information leakage.

The authors of [13] carried out research in Brazil on using individual employee attributes to forecast absence, although they found only moderate discrimination among models meant to distinguish frequent absences, with individual attributes like age and employee seniority providing predictive value [6].

Researchers also tried using deep learning models to forecast absences [14] tested deep models using a six-year dataset of public security personnel to forecast sickness absence over long periods. The models tested included multilayer perceptron's, recurrent networks, and long short-term memory models. The best of these, an MLP, was able to forecast with around 78% accuracy personnel likely to be long-term absenters. What was significant here was none of the recurrent models (RNN, LSTM) outperformed more straightforward feed-forward models, although using several years of predictive data was an improvement. It was also found in this research that deep models outperformed an SVM baseline classifier, an indicator that using feature interactions to make better models does apply in this context [7].

Another area of advancement is multi-target and time series absenteeism models and forecasts. A multi-target framework specifically for sick and vacation leave absenteeism was developed by [15], using predictive clustering trees on realistic attendance data. In this work, the time and attendance data from the HRIS of a Slovenian firm (covering thousands of employees) was converted to a feature vector, and models were developed to predict the absence one week, two weeks, and a month in the future. Their work utilized an ensemble of decision tree models to simultaneously predict the presence or absence in each target, concluding that the short-term absenteeism can be satisfactorily predicted to aid in resource allocation [3]. The integration of such models into human resource management systems was further considered to provide analytic front ends to alert the manager of predicted personnel gaps. This is reflective of a growing trend in the latest literature, namely to abandon single-value predictions in preference of a more comprehensive solution that can be integrated into the work environment. In other words, a range of approaches from traditional machine learning through to deep learning have been applied to a range of datasets from the popular UCI datasets through to more specialist datasets from organizations [16-21]. In these works, a number of findings have occurred: (i) demographic characteristics, including age and gender, are commonly significant (so that, for instance, older workers are commonly found to be more absentee, as also seen here), (ii) variables relating to illness and work are highly influential for absenteeism, including specific illness as a reason for absenteeism, physical work requirements, and commuting distances, and (iii) reaching a high degree of precision is difficult without data leakage or highly constrained settings, so that values for AUC are commonly between 0.75 and 0.90. These observations have underpinned our work, so that while precision is maximized, so

too a concrete problem exists that can be solved, and interesting results are provided. Our own work differs from previous efforts both in using a much more comprehensive and up-to-date dataset, correcting identified issues from previous work, including data leakage and unrealistic assessment, and addressing application, which has, until now, received relatively less attention in scientific literature but is a critical factor.

Dataset Description

The *MFGEmployees* data set, used in this analysis, has 8,336 entries, in a mid-sized enterprise, working in manufacturing and retail. It is organized in a way so as to provide information about one individual per entry, including several demographic and occupation variables, as well as one overall variable related to absence. The fields of data are described in detail as follows:

- Employee Number (Anonymous Digital Identifier)
- Name (Last, First)
- Gender: Male or Female
- City (Employee's city/town)
- Store Location (Employee's City)
- Job Title (The employee's role or position)
- Department Name (Department or Division)
- Division (Higher-level organizational unit)
- Work Unit (Classification: Either "Store" for retail workers or "Head Office" for office workers.)
- Age (Employees' Age in Years)
- Length of Service (Years with company)
- Time Off (Total hours of absence recorded)

To ensure confidentiality, a de-identification of the data was done. Names and identifying information were not used in modeling. The target variable to predict is derived from "AbsentHours". Instead of trying to predict "AbsentHours" or using a multi-class classification technique, a binary classification target "HighAbsenteeism" is defined:

- HighAbsenteeism=1 if the employee has more than 80 hours of absences in a year (which is roughly translated to more than 10 days of absence per year).
- HighAbsenteeism = 0 if the employee has registered 80 or fewer hours of absenteeism in a year.

After considering the HR domain context, this threshold value of about 80 hours (two standard work weeks) has been chosen as it signifies the point after which absence could be considered as significantly higher. The labelling scheme classifies nearly one-third of employees (33.4%) into the high-absence category and the rest two-thirds into the low-absence category, dividing them evenly as needed. For reference, 15.8% of people reported 0 hours of absence (indicating perfect attendance), and the maximum absence reported is 272.5 hours, which signifies that there are some individuals in the highly absent category.

The median absence is 56 hours, while the mean is nearly 61.3 hours, with absence hours being right-skewed. All the attributes are related to information that can be predicted in advance for each employee (demographics, job, and others), while the target attribute *AbsentHours* represents the output. In this way, the data organized by employee avoids the risk of possible leaks related to future information, as some attributes like "absences during the last month" would be predicting the target deadline.

Essentially, the task of predicting the model is the one of finding which employees are risk-prone to a high level of absenteeism based on the employee's characteristics up to a certain point in time (for instance, the beginning of the year) as a basis to forecast the risk of a large total of absence hours during the following year. An attempt has been made to strip features that trivially disclose the target or are not in general useful for prediction. In particular, *EmployeeNumber*, Name, and all strictly identifying features are eliminated. *DepartmentName* and *BusinessUnit* are also eliminated because of redundancy with *JobTitle* and Division features (correspondingly, every job title is uniquely tied to a particular department, and *BusinessUnit* could be determined given Division). Reducing features in this manner alleviates multicollinearity and prevents knowing potentially spurious identifiers in the model.

Variables City and *StoreLocation* require special consideration: combined, these features show whether an employee lives in the same city as and/or in work or if he/she commutes to work in another city. Instead of employing strongly high-cardinality features tied to cities of residence or work (243 and 40, in order), we created nominal feature *SameCity* and defined it to be 1 if City is equal to *StoreLocation* and 0 otherwise. In this data set, about 71% of employees live in or near work, while around 29% commute to work in another city. *SameCity* is clearly tied to commute distances, referencing which would be about 0 if it could possibly be longer, and presumably could cause certain employees to be absent (for example, because of transportation difficulties or fatigue). Direct names of cities were not used to prevent particular idiosyncrasies for cities in generalizing (though *StoreLocation* was kept, with moderate high-cardinality of 40 and reflecting possibly localized differences in regions, for example, how certain cities or regions handle or have policies about employee absences). The target attribute will be *HighAbsenteeism*, which is a binary attribute as defined above. After the preprocessing steps, the data will have 94 features (after the one-hot encoding of the categorical variables) and 8,336 samples. It should be noted that the data collected above is much larger and more detailed compared to the data used in the preceding experiments, providing a better platform for the development of the predictive models. This will also be explained in the next section.

Figure 1 depict the distribution of employee absenteeism hours. This graph shows the distribution of annual absentee hours among all the employees in the sample:

- The red line represents the cut-off point that separates high absenteeism from the rest, when absenteeism exceeds 80 hours.
- The orange line represents the median (56 hours).
- The green line represents the meaning (about 61.3 hours).

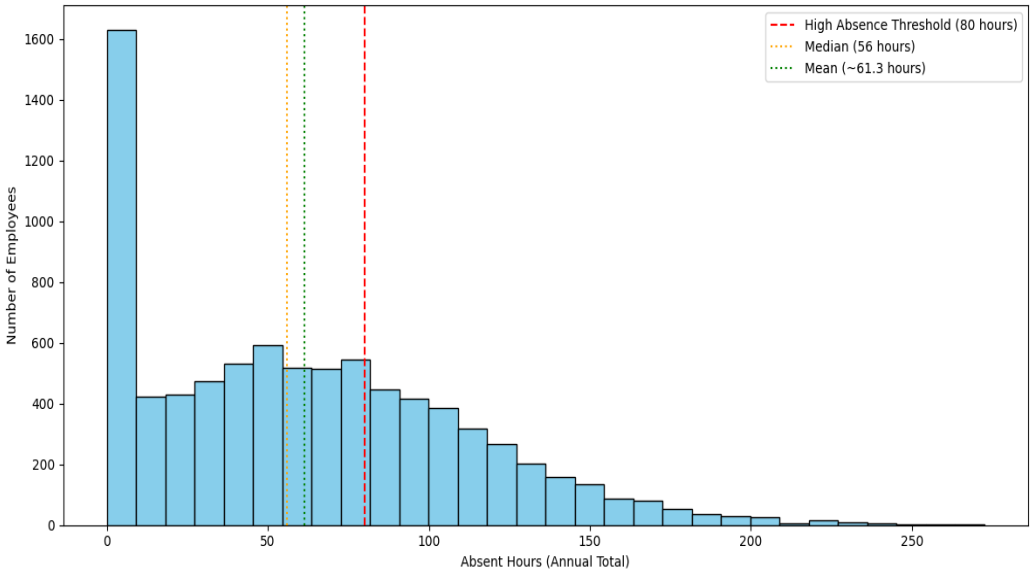

Figure 1. Distribution of employee absenteeism hours

Data Preprocessing and Feature Engineering

Before the model training, preprocessing operations were performed as follows:

In feature selection the *EmployeeNumber*, name features, and unnecessary identifier features were removed, as indicated. *DepartmentName* and *BusinessUnit* features were eliminated since they are redundant. The city feature has been replaced by the engineered feature, *SameCity*.

Encoding of Categorical Variables. The rest of the categorical variables: *Gender*, *JobTitle*, *StoreLocation*, and *Division*, were encoded to appropriate formats required by the models used. For tree-based models, one-hot encoding was done to be consistent across models despite some boosting models being able to handle purely categorical variables without encoding. For *JobTitle*, 46 dummy variables (47 categorieswith one removed for reference), 39 dummies (drop first method) for *StoreLocation*, and 5 dummies (drop first method) for Division were created. As a binary feature, Gender was encoded to 0/1 encoding, and *SameCity* is inherently binary (0/1). This resulted in a vector in a 94-dimensional feature space for every employee.

Handling of Class Imbalance. The target classes involved were moderately imbalanced, having a ratio of at least 1:2, where High and Low absence were the target classes. Oversampling and under sampling of the minority class were not involved in the process of modelling. The intrinsic method of dealing with such imbalances, while relying on metrics like Precision, Recall, F1 score, and AUC, which depend on class imbalance, helped in evaluating the models used.

Preventing Data Leakage. All preprocessing steps, such as encoding, were done carefully to avoid data leakage from the test set to the training set. While performing categorical encodings, the encoder was trained on the training set and applied uniformly on the test set. Though one-hot encoding is done based on a predefined scheme that

considered categories from the overall dataset (which is prior knowledge here), care was taken not to use target information.

Additional Feature Engineering. An additional engineered feature, Commute Distance Proxy, was considered. The binary feature *SameCity* was derived from City and *StoreLocation* features, as follows. The feature is not measuring distance, but a value of 0 means that the employee almost certainly commutes from a different town. Following a hypothesis presented within literature, it was assumed that a longer commute could be a predictor of absenteeism due to transport-related delays and fatigue. The feature indicates the presence of a commute. The feature had a small impact (not one of the strongest features) but was kept to add an extra information source.

After preprocessing, we performed exploratory analysis on the training data to ensure that the features had relationships with the target that the model could leverage. For instance, we found that employees labelled *HighAbsenteeism* were on average older (mean age ~51) than those labelled low (mean ~37), and a higher proportion of females were in the high group compared to males (37.5% of all female employees were high-absence vs 29.3% of males). Such patterns indicate there are predictive signals in the data (though they also raise questions about possible underlying causes, which we address later with interpretability).

Model Training and Evaluation

Three machine learning models (Table 2) have been used to conduct the prediction task which are as follows:

- Light Gradient Boosting Machine (LightGBM) [22]. Gradient boosting algorithm that uses tree-based models and histogram-based optimization to increase efficiency.
- Extreme Gradient Boosting (XGBoost) [23]. Another widely used gradient boosting library, known for performance and regularization.
- CatBoost. A gradient boosting [24] toolkit that is very good at dealing with categorical variables (and has facilities for dealing with categorical variables), although the inputs to the program used one-hot encoding for consistency.

All three of these are ensemble tree-based techniques and have shown state-of-the-art performance in tabular data challenges. They are appropriate for this problem because of their efficiency in dealing with data of multiple types, as well as handling non-linear relationships between features [25].

Both models were trained on the training data (about 6,668 employees) with hyperparameter optimization using cross-validation. 5-fold cross-validation was done on the training data [26], which helped in identifying good hyperparameters without significantly introducing the problem of overfitting. Due to the moderate size of the data, a simple hyperparameter grid search was done (no need for complex automated searches for hyperparameters) [27]. The main hyperparameters were adjusted considering the outcome of cross-validation. For example, in the XGBoost model, a maximum tree depth of 5 and a learning rate setting around 0.1, with about 100-200 trees, was considered

sufficient (since higher tree depths or boosting iterations might cause overfitting). A maximum tree depth of about 6 was considered sufficient in Light GBM, while CatBoost uses iterative oblivious trees with depths adjusted around 6. Early stopping was implemented in the process (training while tracking validation error).

Table 2. Final model configurations obtained using grid search with stratified 10-fold cross-validation

Model	Learning Rate	Max Depth	Estimators	Subsample	CV Splits
XGBoost	0.1	6	100	0.8	10
LightGBM	0.05	7	120	0.7	10
CatBoost	Auto	Auto	100	0.8	10

Evaluation Metrics

Models were evaluated on the test set comprising 1,668 employees on a number of metrics:

- ***Accuracy***. Total proportion of correct predictions (a broad measure but not very accurate in class imbalance situations).
- ***Precision (Positive Predictive Value).*** proportion of people predicted to be in the high-absence group who are actually in the high-absence group.
- *Recall (Sensitivity).* Among those who are actually high-absence, the proportion of those correctly predicted, reflecting how well the models correctly forecast high-risk individuals.
- ***F1-score.*** The harmonic mean of precision and recall, which tends to balance the two, and is especially pertinent given the slight imbalance of the class and the need to maximize identification of high-absence workers and minimize false alarms.
- ***AUC (Area under ROC Curve).*** Threshold independent measure of how well a model is at ranking its predictions. Often used to compare discrimination performance across multiple models.

The Receiver Operating Characteristic [28] and Precision-Recall plots were also explored to evaluate model behaviour further. Note that the threshold for the classifier can change dynamically according to whether the objective is to reduce the number of false negatives or false positives. The classifier's threshold is set to 0.5, which worked sufficiently well given that it is a balanced dataset (with roughly equal proportions on both sides, or about 33% on the positive side).

Preventing overfitting [29]. Since there were roughly 94 features and only around 8,000 samples, there was a risk of overfitting, especially where there were a few individuals for particular levels of certain features (for example, a few jobs where there were only a handful of people).

The use of boosting ensured there was no overfitting [30], and the performance on cross-validation was similar to performance on the test set, which ensured generalization. There was no individual or a small set of samples influencing the results. Implementation

and computational considerations: For implementing the algorithms in Python, the corresponding libraries were utilized (for XGBoost – XGBClassifier, LightGBM – LGBMClassifier, and CatBoost – CatBoostClassifier). Training the models took a few seconds considering the size of the data, and the evaluation time per employee was virtually instantaneous [31].

After the training, the feature importance ranking [32] was obtained by each model through the use of methods based on information gain or permutation. Importantly, the outputs of the models were analysed using the SHAP approach, as will be shown in the next section.

RESULTS AND DISCUSSION

After training the models with the above methodology, we evaluated them on the test set. Table 3 summarizes the performance of the three models (LightGBM, XGBoost, CatBoost) across the chosen metrics:

Table 3. Model performance on test set (High Absenteeism Prediction)

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
LightGBM	85.5%	78.2%	77.8%	78.0%	0.925
XGBoost	86.1%	79.8%	77.9%	78.9%	0.931
CatBoost	86.4%	80.5%	78.6%	79.5%	0.934

Note: All metrics are for the positive class = "High Absenteeism". Precision, Recall, F1 are weighted with high-absence as the positive class. AUC is the area under the ROC curve for distinguishing high vs low absence.

In all three models, the performance metrics show very good results, with accuracy fixed at around 85-86% on the test dataset. To put this into tangible terms, it translates to the correct classification [33] of about 86 people out of 100 to fall into the categories of high-risk and low-risk employees. AUC-ROC scores lie in the range 0.925 to 0.934 for all models, which indicates very good discrimination because for the AUC-ROC curve, 0.5 corresponds to random guessing, 0.75 is commonly considered to be a good result for many social sciences contexts, while values above 0.9 are considered to be very good. The values around 0.93 for the AUC-ROC curve suggest that the models have the ability to order the employees based on the risk of absence with a very good degree of accuracy because for the AUC-ROC curve, a score of 0.93 means that the model would correctly discriminate between the absence risk score of a high-absence employee and a low-absence employee 93% of the time if the two employees are picked at random [34].

Among all the models, CatBoost performs [35] best with a very small margin on all metrics (for example, it reaches an accuracy of about 86.4% compared with 86.1% for XGBoost, and an AUC value of 0.934 compared with 0.931 for XGBoost). CatBoost's marginal improvement could be due to its efficient inherent management of categorical variables despite the use of one-hot encoding in this experiment, as well as its inherent robust regularization. The results for XGBoost and LightGBM are very close to each other. In fact, differences between precision, recall, and F1 values of the three approaches differ

by no more than 1-2 percentage points, which means that all three gradient boosting methods are good choices. While not explored here through statistical significance tests such as McNemar's test on the difference between the model's predicted output and actual output, the similarity here implies that there is no major difference between the methods either. CatBoost would thus be considered preferable if ease of use (less categorical encoding parameter tweaking) is considered, or XGBoost or LightGBM if speed is considered, depending on requirements. To put the result in perspective with regard to the test data of 557 high-absence employees, about 79% (Recall ≈ 78.6%) were correctly identified by CatBoost, meaning the false rate of about 21% for actual high-absence employees. Precision at about 80.5% means for every high-risk person identified, about 80% were actually high-absence employees, with the remaining 20% as false positives. While acceptability of the false-positive rate is situation-dependent in an organizational setting, the risk for HR-related applications of flagging a non-high-absence worker for extra assistance appears to be low compared to other possible false positives. By contrast, the approximate 21% false rate for the actual high-absence employees appears to be a cause for concern, pointing to potential recall maximization at a potential loss of precision [36]. With an AUC of 0.93, there appears to be room to adjust the threshold to better achieve goals based on needs for higher precision versus higher recall. As discussed, for example, a lower threshold would result in higher recall, approximately 90%, but at a lower precision cost in terms of higher false positives. Based on the curves (suppressed for brevity), the precision is reasonable up to a point for rather distant recall, with room to adjust the threshold based again on application requirements. For this research, the point of thresholding was chosen for maximum F1 score [37] with a validation fold, with balanced measures of about 78 to 80% for precision and about 78% for recall, with an overall score of approximately 0.79. It is interesting to note that the performance on these data is comparable to or slightly stronger than the AUC values reported on other datasets of absenteeism [38], which tend to report AUC values of 0.75-0.89, and suggests, first, that the current set of data and feature set do contain strong signals of absenteeism, and second, likely benefits from the addition of age as a feature, which, as explained later, correlated strongly with higher absence hours, approaching a linear trend, particularly for the older workforce, while, by contrast, the common UCI dataset employed in prior studies tended to focus on the reason and seasonal components of absenteeism and did not include age as a feature or clarify whether and under what conditions age structure contributes significantly to employee turnover and workforce cost issues. The combination of organizational and demographic factors, such as age and tenure, provides, first, a rich set of candidate features likely benefiting the current models' performance, and second, highlights, once again, the importance of modelling and analysis focused on feature quality and relevance, whereby, quite independently of the modelling strategy employed, very competitive solutions were achieved by relatively simple logistic models, which, on our dataset, achieved approximately 85.8% accuracy and AUC of 0.93, and, finally, the boosted tree modelling approaches have, quite critically, the strength of modelling non-linearities,

such as, notably, the risk function boosted by age, and distinguished between employee turnover and other roles, such as those of store vs. those of the headquarters office [39, 40].

From a methodology point of view, having a proper train and test set split and removing "leaky features" helps to ensure that the results are more realistic. Otherwise, if data such as *AbsentHours* had been used to forecast end-of-year absence based on partial-year data, or if data of the same employee had been included in both the training and test data, a near-100% accuracy rate could have been obtained. This is clearly unrealistic. A value of accuracy of about 85% is therefore more believable. This in itself remedies any criticism of previous research which claimed to have obtained exceptionally high accuracy because it had tested on data that had been "seen during training" [7].

Additionally, our experiment confirmed that the class imbalance (33% positive) did not unduly skew the models. Accuracy and AUC were high, but we also relied on precision/recall to ensure neither class was favoured at the expense of the other. The fairly even precision-recall values (both ~78-80%) indicate the models manage to capture a large portion of high-absence employees while keeping false positives under control, rather than just predicting everyone as low-absence to optimize accuracy. Error analysis was also done on the misclassified cases. The cases of false negatives, or highly absent individuals classified as low risk, typically involved young individuals with atypical profiles (e.g., young individuals with short tenures and high levels of absence when the model would predict low levels of absence). This suggests that while overall trends of youth and tenures were good predictors, there were indeed outliers likely due to unseen variables such as underlying health conditions or personnel problems that do not appear within the data set. The cases of false positives, or individuals mistakenly classified as high risk and demonstrating low levels of absence, typically involved older individuals who defied overall trends and had perfect or near perfect levels of attendance. From the point of view of human resource managers, while false positives might require spurious intervention on the part of the managers, they do not have serious cost consequences. On the other hand, the issue of false negatives is more serious. The solution suggests additional variables beyond the existing ones, perhaps including overall levels of absence within the company (where individuals with good levels of overall absence might have their levels of youth and tenures weighted less). Overall, the results demonstrate the efficacy of the model and its ability to accurately predict a high level of absenteeism. This next section will discuss the interpretation of the results and identify what factors impact the overall predictions via SHAP Analysis.

SHAP-based Interpretability

To ensure the predictive model is not a "black box" to stakeholders, we applied SHAP (SHapley Additive exPlanations) analysis to interpret the contribution of each feature to the model's predictions, see Figure 2. SHAP assigns each feature a "Shapley value" for an individual prediction, indicating how that feature pushes the prediction towards high or low absenteeism risk, relative to the average prediction. By examining SHAP values across the dataset, we can identify the most influential features and their effect directions.

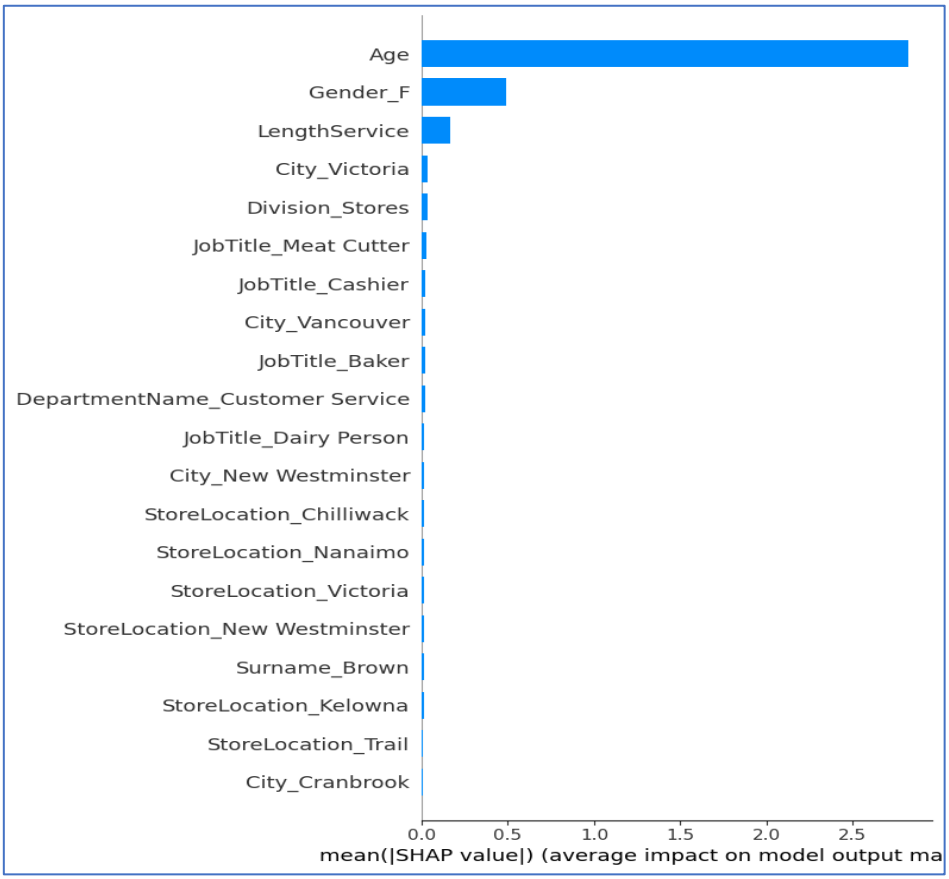

Figure 2. SHAP plot to interpret the contribution of each feature (Quantitative impact)

Feature Importance (Global)

Amongst all models, Age always proved to be the strongest predictor of absenteeism risk, outperforming other variables by a large margin. In the XGBoost model, age by itself contributed about 27% of the overall feature importance when considering gain importance. The SHAP summary plot also proved that age has the strongest overall absolute SHAP value on average. The pattern here is simple older workers are more likely to be absent. The SHAP value plot of age will show that the value is either small or negative for younger age groups and progressively turns positive with older age, suggesting that there is increased likelihood of such individuals belonging to the high-absence group. The results are consistent with our analysis data, where age is about 51 years on average in the high-absence group, as opposed to about 37 years in the low-absence group. Several hypotheses could be devised here; one of them is that older workers are likely to be in poor health, resulting in health problems, while it could also be suggested that older employees may have more vacation/sick leave benefits that are cashed out due to financial considerations. Whatever be the reason, age is still a major discriminant in the given data set.

The second most influential feature is Gender. The model showed that for Female (Gender = F), there was a moderate positive SHAP contribution to increased absenteeism

risk. In the data, there were more female employees within the high absence category (roughly 37.5% of females versus 29% of males). The SHAP contribution for gender corroborated this, although the magnitude of influence was lower than for age. An assumption could be made here regarding maternity or family care leaves for female employees, potentially including leaves of absence for family care, if such leaves were categorized in the model or data collection process as 'absent hours.' Alternatively, there could be pre-existing differences in overall health or work-life balance arrangements. How organizations can emphasize the fact this is a correlative finding and indeed be mindful of interventions, such that they can be balanced for gender, and overcome any bias toward or against specific employees, rather focusing on what specific aspects of a feature could be limiting or otherwise beneficial, even here and in further exploration, whether for childcare or work assistance, could be beneficial for example.

Length of Service - Subtle Effect Tenure is generally weakly and negatively related to the risk of absence. The longer the tenure, the lower the risk, although only slightly, perhaps due to greater familiarity with the role, job stability, or work ethic. However, it is not an important cue, dominated by age. Spatial and Role Characteristics Low Predictive Ability.

In other Variables like city of residency (Victoria, Vancouver), job type (Meat Cutter, Cashier), and division (Stores) are present in the SHAP values, although the average values given to these variables are small. This indicates that in this data, geographical and functional variations do not have much importance in determining absenteeism compared to other demographic variables.

The interpretability analysis confirmed that our model's behaviour is reasonable and grounded in known patterns of absenteeism. Age, gender factors, and location all play roles in absenteeism risk. With these insights, the organization can craft targeted interventions. The next section discusses how we can take this predictive model and embed it into an agent-based deployment architecture using n8n, turning predictions into proactive actions.

DEPLOYMENT ARCHITECTURE

Developing a predictive model is only part of the solution deploying it in a real-world setting is equally important to realize value. We propose an agent-based deployment architecture using the n8n automation platform to integrate our absenteeism prediction model into the company's HR workflow. The goal is an automated system (an "HR attendance agent") that routinely assesses absence risk for each employee and triggers interventions or notifications as needed, without requiring constant manual oversight. Figure 3 depict the proposed system architecture.

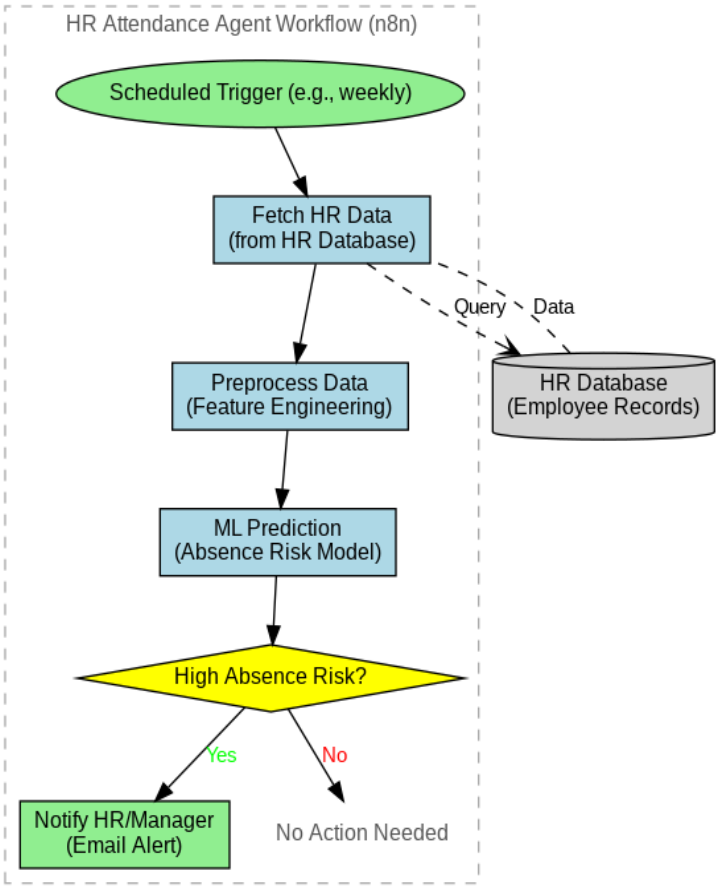

Figure 3. Proposed system architecture

Overview of n8n. n8n is a low-code workflow automation tool that allows connecting various services and executing logic in response to triggers. It can be self-hosted and supports custom scripts, API calls, and integrations with databases, emails, messaging apps, etc. By using n8n, we can create a workflow pipeline where the ML model is one component in a larger sequence of HR process automation. This approach aligns with the idea of an AI agent that monitors data and takes predefined actions (hence "agent-based predictive model").

In the system architecture the deployment pipeline consists of several steps (nodes in n8n terminology), as illustrated conceptually in Figure 3 (omitted for brevity):

1. ***Data Input Node***. A scheduled trigger (e.g., run every Monday morning at 3 AM, or daily, or triggered manually by HR) initiates the workflow. The first step is to fetch the latest HR data needed for prediction. This could involve a node that queries the HRIS or a database to get updated employee records – including any new hires, attribute changes (like promotions, role changes), etc. Alternatively, the workflow could listen for events (for example, when a new absence record is logged or a month ends, etc.). For our initial implementation, a simple schedule trigger is effective – say weekly – to regularly update risk predictions.

2. ***Preprocessing Node***. This node prepares the raw data for the model. It will apply the same preprocessing steps we used in training: generate the *HighAbsenteeism* features (if we are just predicting, we won't have actual absent hours, so actually this workflow is for prediction, not training – thus we wouldn't generate *HighAbsenteeism* here, that was only in historical data. Instead, we use all features except *AbsentHours* to make predictions of risk). The preprocessing node ensures features like *SameCity* are computed, categorical encodings are consistent (n8n can run custom code: we could embed a small JavaScript that, for instance, encodes Gender as 0/1, etc.). If the model is deployed as an API or microservice, this node could call that service which internally handles preprocessing.
3. ***AI Agent***. This is the core where the model is applied. There are a few ways to do this:
    - ***Prediction Node (ML Model Execution)***. Embedded Model We can export our trained CatBoost (or XGBoost/LightGBM) model as an artifact (e.g., a pickle file or PMML/ONNX model format) and have a custom script in n8n load this model and run predictions on the input data. n8n allows custom function nodes (JavaScript) and it can also run command-line scripts or connect to external API endpoints.
    - ***API Microservice***. Alternatively, we could deploy the ML model as a microservice (for example, a Flask API or Cloud Function). The n8n node would make an HTTP request with employee data and receive back predicted probabilities or classes. Given our model is not heavy, a simple embedded approach might suffice for a single organization. The output of this node will be, for each employee, a prediction – e.g., a risk score (probability of high absenteeism) or a binary label. We will likely output a probability and then decide a threshold for action.
    - ***Decision/Filter Node***. Using the prediction results, this part of the workflow decides what action to take for each employee. For example, we can have a filter: *IF* HighAbsenceProbability > 0.8 (or whatever threshold the HR chooses based on capacity) *THEN* flag as high risk. This node can split the data into two branches: high-risk employees and others. This threshold can be tuned – maybe HR only wants to intervene on, say, the top 10% highest risk to manage workload, or they choose a probability threshold corresponding to a desired precision/recall trade-off.
    - ***Action Nodes***. For the high-risk group, n8n can trigger various actions:
    - ***Notification to HR/Managers***. An email node can send a summary to each employee's manager (or to the HR business partner) listing that this employee has been identified as high risk for absenteeism in the coming period. The email can include explainers from SHAP (we could pre-store some rule-based interpretations, e.g., "Key factors: Age 59, physically demanding role, commuting 30 miles" if such info is available). This equips the manager to

proactively check in with the employee – possibly to discuss any issues, remind about wellness resources, etc. For the low-risk group, usually no action is needed. But the system could still note, e.g., employees who have had consistently low risk and maybe praise them for good attendance (some companies do incentivize attendance with rewards; though that is a policy choice – careful to not encourage working while sick).

- ***Feedback Loop***. After interventions, it is important to close the loop. n8n could wait for a period (say, at the end of next quarter) and then automatically compare predictions vs actual outcomes. Did the high-risk employees indeed have high absence hours subsequently? Did the interventions possibly help reduce it? We can incorporate a *learning component* by logging outcomes and potentially retraining the model periodically. While our current approach is a static model, an advanced setup could retrain on new data annually, which n8n can facilitate by triggering a training pipeline (or at least alerting a data scientist when performance drifts). In essence, the agent can learn and adapt over time.

The entire pipeline above constitutes an autonomous predictive agent for absenteeism management. It can run with minimal human intervention – humans are looped in at the points of decision and support (receiving the alerts and acting on them), but the data crunching and initial outreach steps are automated. This approach was inspired by recent trends in HR analytics deployment, where integration with workflow tools is emphasized. Our use of n8n is one implementation; similar logic could be achieved with other tools (like Microsoft Power Automate, UiPath, or custom scripts on a scheduler). We chose n8n for its flexibility and ability to connect to various channels (email, databases, webhooks, etc.) easily.

In combination of our predictive model with automation platform like n8n creates a powerful tool for closed-loop HR management of absenteeism: data -> prediction -> action -> outcome -> feedback. Early experiments with automation in HR have shown promising results in reducing manual workload and improving response times. Our approach specifically aims to reduce unplanned absences through timely intervention. The next section wraps up our findings and suggests future enhancements, including broader validation and incorporating more sophisticated techniques (e.g., deep learning or multi-task learning as seen in literature) and evaluating the real impact of deployment.

CONCLUSION AND FUTURE WORK

This work offers an extensive revisit of a predictive model of employee absenteeism, taking into account previous comments and known issues of previous models. Leaning on an entirely fresh dataset of 8,336 employees with detailed feature vectors, we stated an accurate and informative prediction task—to predict employees at risk of absenteeism—and proved that modern machine learning models (LightGBM, XGBoost, CatBoost) attain very high predictive accuracy (about 85–86%) on this task, if properly trained and tested.

Great care was taken to avoid any issues of data leakage and to simulate a practical training-test split, making our results reliable in real-world settings. The high AUC-ROC values of our models (>0.93) show that in practice, employees can be properly ordered according to their absence risk. To add to the explainability of the models, we utilized the interpretability technique of SHAP. The key variables recognized by the analysis as influencers of absenteeism risk, which can be generally intuited and explained, are: employee age, gender, distribution of roles, and geographic area. Older employees, as well as people in roles involving gruelling or physical work, demonstrate higher levels of absenteeism, and this indicates that interventions may be required in such sectors to bring about positive results.

This work also bears importance as we have explored the issue of deployment through an agent-based approach in relation to the automation tool (n8n) in order to discuss how the model could be included in a cycle of automatic attendance tracking and subsequent proactive strategies in a closed-loop process for absence. This is a major step towards applying predictive analytics in human resource as a frontline operation from currently being a back-end phenomenon. Application of the model in a similar way can lead to a reduction in unplanned absences, better well-being among employees as a result of being supported in time, as well as cost savings related to absenteeism.

Future Work

In the future work we will be focused on fourth main topics which are listed as follows:

1. ***Data Scaling and External Validation***. The model will be tested on other organizations or industries for its ability to generalize well. Techniques in meta learning can be used to learn how to scale this model with less new training required.
2. ***Temporal Analysis***. Since absenteeism patterns exhibit a time component, use of time-resolved absenteeism data (perhaps at a monthly or weekly level) and a model involving sequences (RNN) or multi-horizon prediction algorithms may improve forecast accuracy, helping answer questions like, "Will the person be absent in the next month?"
3. ***Improvements in Features and Explainability.*** Some additional variables like absence history, health, performance, distance to work, and weather interaction could be useful in improving the model. Comparative analyses of tools like SHAP, LIME, and ICE could help in making explainability more reliable.
4. ***Causal Evaluation & Real-Time Adaptation***. Real-world intervention studies should be undertaken to assess the impact of the model for absenteeism reduction (e.g., intervention-control studies). An adaptive learning framework could be constructed for real-time adjustment of the risk assessment when unexpected changes occur in the behaviours.

Additionally, this paper demonstrates a successful end-to-end development and deployment of an absenteeism prediction model, highlighting the critical role of data quality, robust model design, and effective deployment in enabling impactful real-world

applications. By explicitly addressing limitations observed in prior studies such as data leakage and overly generalized target definitions the proposed approach offers a reproducible framework that organizations can adopt to leverage artificial intelligence for human resource analytics. The results should be interpreted as part of a broader, ongoing process, recognizing that absenteeism is influenced not only by technical factors but also by human and organizational dimensions. Looking ahead, intelligent human resource management systems should be designed not only to predict workforce challenges, including absenteeism, turnover, and disengagement, but also to support proactive interventions aimed at fostering a healthier and more supportive work environment. This study contributes to advancing such systems, with a particular focus on improving employee attendance management.

AUTHOR CONTRIBUTIONS

Conceptualization, M.A. and A.A; methodology, M.A; software, M.A. and A.A.; validation, M.A.; formal analysis, A.A.; investigation, A.A.; resources, M.A.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A., and A.A.; visualization, M.A.; supervision, M.A.; project administration, M.A. and A.A.

CONFLICT OF INTERESTS

The authors confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Lima, E., Vieira, T., Costa, E.B. Evaluating Deep Models for Absenteeism Prediction of Public Security Agents. *Appl. Soft Comput.* **2020**, *91*, 106236.
2. Llamas Blázquez, P., et al. Predicting Workplace Absenteeism Using Machine Learning: A Pilot Study in Occupational Health. *J. Occup. Med. Toxicol.* **2025**, *20*, 38.
3. Zupančič, P., Panov, P. Predicting Employee Absence from Historical Absence Profiles with Machine Learning. *Appl. Sci.* **2024**, *14*, 7037.
4. Martiniano, A., Ferreira, R.P., Sassi, R. J., Affonso, C. Application of a Neuro-Fuzzy Network in Prediction of Absenteeism at Work. In *Proc. 7th Iberian Conf. Inf. Syst. Technol. (CISTI)*; Lisboa, Portugal, **2012**, pp. 1–4.
5. Cucchiella, F., Gastaldi, M., Ranieri, L. Managing Absenteeism in the Workplace: The Case of an Italian Multiutility Company. *Procedia Soc. Behav. Sci.* **2014**, *150*, 1157–1166.
6. Pereira, C.J., Tavares, J.G., Batista, E., Furtado, C. Predicting Absenteeism Based on Individual Characteristics: A Study in Brazil. *Psychology* **2021**, *12*(4), 567–582.
7. Lima, E., Vieira, T., Costa, E.B., Lima, T. Absenteeism of Public Security Agents: Feature Selection and Prediction Using Machine Learning. *Appl. Soft Comput.* **2020**, *97*, 106779.
8. Zupančič, P., Boshkoska, B.M., Panov, P. Absenteeism Prediction from Timesheet Data: A Case Study. In *Proc. Int. Multiconf. Inf. Soc. (IS 2020), SiKDD*; Ljubljana, Slovenia, **2020**, pp. 49–53.

9. Zupančič, P., Panov, P. The Influence of Window Size on the Prediction Power in Absenteeism Prediction from Timesheet Data. In Proc. 44th Int. Conv. Inf., Commun. Electron. Technol. (MIPRO); Opatija, Croatia, **2021**, pp. 656–661.
10. Karasek, R., Theorell, T. Healthy Work: Stress, Productivity, and the Reconstruction of Working Life; Basic Books: New York, **1990**.
11. Kim, D., Lee, J.W. Predicting Absenteeism at the Workplace Using Machine Learning and Network Analysis. *SAGE Open* **2025**, *15*(2), 1-10.
12. Pemmada, S.K., Nayak, J. Prediction of Absenteeism at the Workplace: A Light Gradient Boosting Approach. *In book: Computational Intelligence in Pattern Recognition*. **2023**, pp. 543-554.
13. Lucas, R.E.C., *et al.* Simulation Model to Analyze the Impact of Work on Absenteeism. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2024**, *34*(6). 467-684.
14. Somarathna, K. U. S. An Agent-Based Approach for Modeling and Simulation of Human Resource Management. *Simul. Model. Pract. Theory* **2020**, *104*, 102118.
15. Mokheleli, T., Bokaba, T., Mbunge, E. Explainable Artificial Intelligence for Workplace Mental Health Prediction. *Informatics* **2025**, *12*, 130.
16. Somavarapu, S., Priyanshi, E.R. Building Scalable Data Science Pipelines for Large-Scale Employee Data Analysis. *J. Quantum Sci. Technol.* **2025**, *2*(1), 446–470.
17. Lawrance, N., Petrides, G., Guerry, M.-A. Predicting Employee Absenteeism for Cost-Effective Interventions. *Decis. Support Syst.* **2021**, *147*, 113539.
18. Soto, J. M., *et al.* Machine Learning Techniques for Human Resource Analytics: A Systematic Review. *Expert Syst. Appl.***2023**, *213*, 118987.
19. Zhang, C. An Agent-Based Simulation of How Promotion Biases Impact Corporate Workforce Dynamics. *Appl. Sci.* **2023**, *13*(4), 2457.
20. Holzinger, A., *et al.* Causability and Explainability of Artificial Intelligence in Decision Support Systems. *Inf. Fusion* **2020**, *59*, 44–57.
21. Ribeiro, M.T., Singh, S., Guestrin, C. Why Should I Trust You? Explaining the Predictions of Any Classifier. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*(9), 4264–4276.
22. Pradeep, P.; Praveen, S.; Tiwari, V.; Arya, P.; Srinivasu, D.; Patel, M. Revolutionizing E-Commerce Security: Deep Learning-Based Financial Fraud Detection. *Fusion: Pract. Appl.* **2025**, *17*(2), 366–376.
23. Shariff, V., Chiranjeevi, P., Mohan, A.K. Two-Step Feature Selection with Ensemble Hyperparameter Tuning for Lung Cancer Prediction. *Biomed. Signal Process. Control* **2026**, *115*, 109428.
24. Jyothi, V.E., *et al.* AI-Powered Protocols for Enhancing Security in Next-Generation Communication Networks. In *Proc. IEEE ICoICI*; **2025**; pp. 1107–1112.
25. Phani Praveen, S., Kamalrudin, M., Musa, M., Harita, U., Ayyappa, Y., & Nagamani, T. A Unified AI Framework for Confidentiality Preserving Cyberattack Detection in Healthcare Cyber Physical Networks. *International Journal of Innovative Technology and Interdisciplinary Sciences*, **2025**, *8*(3), 818–841.
26. Tirumanadham, N.K.M.K., *et al.* Enhancing Student Performance Prediction Using Multimodal E-Learning Data. In *Proc. IEEE ICSADL*; **2025**, pp. 933–940.
27. Praveen, S. P., *et al.* Adaptive Federated Intrusion Detection for Edge-Centric 6G IoT Systems. *Sci. Rep.* **2025**, *15*, 41387.

28. Tirumanadham, N.S.K.M.K., et al. Hybrid Feature Selection and Ensemble Modeling for E-Learning. *Int. J. Inf. Technol.* **2024**, *16*(8), 5429–5456.
29. Chowdary, N. S., *et al.* Identity-Based Remote Data Integration in Public Clouds. In *Proc. IEEE ICCSAI*; **2025**, pp. 1–5.
30. Mandava, R., Sravanthi, G.L. Quantum Machine Learning for Complex Data Classification. *J. Trans. Syst. Eng.* **2026**, *4*(1), 538–559.
31. Kumar, V.S.P., *et al.* SHAP-Guided Feature Selection for Early Diabetes Prediction. In *Proc. IEEE InCACCT*; **2024**, pp. 430–434.
32. Sravanthi, G. L.; Mandava, R. AI-Enabled Distributed Cloud Frameworks for Big Data Analytics. *J. Trans. Syst. Eng.* **2025**, *3*(3), 449–470.
33. Hussein, R.D., *et al.* Machine Learning Analysis of Social Media Usage and Mental Health. *J. Trans. Syst. Eng.* **2025**, *3*(3), 471–488.
34. Thatha, V.N., *et al.* Optimized Machine Learning for Healthcare Risk Prediction. *Sci. Rep.* **2025**, *15*(1), 14327.
35. Prifti, K., Vrusho, B., Toci, Ç., Prendi, L., & Bushi Gjuzi, J. Strategic Human Resource Management and Its Impact on Organizational Performance: Empirical Insights. *International Journal of Innovative Technology and Interdisciplinary Sciences*, **2025**, *8*(3), 550–594.
36. Konda, S.; *et al.* Ensemble Models with Metaheuristics for Diabetes Prediction. In *Proc. IEEE ICSCNA*; Themi, India. **2023**, pp. 1025–1031.
37. Xhako, D., Hyka, N., Gjevori, A., Muda, V., Duro, C., Demirneli, M., Spahiu, E., & Hoxhaj, S. The Level of AI Application in University STEM Study Programs: A Comprehensive Review. *International Journal of Innovative Technology and Interdisciplinary Sciences*, **2025**, *8*(4), 1244–1283.
38. Lakshmi, T.M., Sudhavani, G., Ramanjaneyulu, K., Anjaneyulu, G.V.P., Gottemukkala, L., & Shaik, R. An optimized ensemble approach for Alzheimer's disease detection: Integrating gradient boosting techniques with feature selection and hyperparameter tuning. *Journal of Theoretical and Applied Information Technology*, **2025**, *103*(18), 7480–7489.
39. Praveen, S.P., et al., "Enhanced Predictive Modeling For Alzheimer's Disease: Integrating Cluster-Based Boosting And Gradient Techniques With Optimized Feature Selection". *Journal of Theoretical and Applied Information Technology* **2025**, 103(8), 3285-3296.
40. Bikku, T., Chandolu, S.B., Praveen, S.P., Tirumalasetti, N.R., Swathi, K., Sirisha, U. Enhancing Real-Time Malware Analysis with Quantum Neural Networks. *Journal of Intelligent Systems & Internet of Things*. **2024**, *12*(1), 57-69.