

PREDICTING EMPLOYEE ATTRITION AND PERFORMANCE USING DEEP LEARNING

SAMER M. ARQAWI¹, MOHAMMED A. ABU RUMMAN², EMAN AKEF ZITAWI³, ANEES HUSNI RABAYA⁴, AHMAD SALEH SADAQA⁵, BASEM S. ABUNASSER⁶, SAMY S. ABU-NASER⁷

¹Associate Professor at Industrial Management Department, Palestine Technical University-Kadoorie, Palestine

²Associate Professor, (MBA, Ph.D), Business Administration Department, Faculty of Business, Al- Balqa' Applied University, Jordan

³Researcher in educational administration, College of Graduate Studies - Department of Educational Administration, Arab American University Palestine.

⁴Head of the Health Administration Department, Al Quds Open University, Palestine

⁵Assistant professor, Arab American University, Higher Education College, Strategic planning programme

⁶University Malaysia of Computer Science & Engineering (UNIMY), Cyberjaya, Malaysia

⁷Professor of Data Science, Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine

E-mail: ¹s.arqawi@ptuk.edu.ps, ²aburumman@bau.edu.jo, ³Eman.sb2017@gmail.com, ⁴anrabya@qou.edu, ⁵ahmad202@yahoo.com, ⁶p05210002@student.unimy.edu, ⁷abunaser@alazhar.edu.ps

ABSTRACT

Making decision can have a vital role in the administration and might indicate the most significant constituent in the route of planning. Attrition of employees is a well-known issue that requires the correct judgments from the management to keep highly skilled employees. Excitingly, Artificial Intelligence (AI), Machine and Deep Learning were applied broadly for instance like an effective means for the prediction of employee attrition. The aim of this study was to utilize machine and deep learning models to predict employee attrition with a high accuracy; furthermore, to identify the most influential factors affecting employee attrition. The dataset used in this study was collected from Kaggle Depository. The dataset was created by the IBM analytics that consists of 35 features from 1,470 employees. To get the best accuracy of prediction of employee attrition, we preprocessed the dataset, balanced it and split it into three sets: train, valid, and test datasets. Several experiments were carried out to show the practical value of this study. The deep learning model archived f1-score of (94.52%), recall (94.52%), and precision (94.58), accuracy (94.52%), whereas the best machine learning model archived f1-score (92.52%), recall (92.55%), precision (92.52), and accuracy (92.55%) for the prediction of employee attrition.

Keywords: *Employee Attrition; Machine Learning; Deep learning; Prediction*

1. INTRODUCTION

The competition between enterprises and companies extremely dependent on the throughput of the employees. Constructing and preserving an appropriate atmosphere is the main key for the contributions for steady and cooperative employees. The division of Human Resource (HR) must contribute in constructing an atmosphere via

investigating history of employees. Carrying out the analysis of these dataset permits the management to enhance the decision determination to elude employee attrition [1]. Attrition of an employee is defined as the techniques that innovative employees choose departing the business as a result of various causes for instance exertion effort, inappropriate setting, or insufficient salary. Attrition of an employee disturbs the company's throughput since it

is going to lose a creative worker in addition to other means like HR team exertion in employing fresh staffs [2]. Hiring fresh workers needs preparation, improvement, and incorporating the workers into a different setting. The prediction of attrition of employee early afore it happens can assist to decrease its influence or the management to stop it. The collected studies advised that pleased and enthused workers have a tendency to be creative, extra innovative and achieve much better [2]. Companies can employ their HR dataset to perform this predictions dependent machine and deep learning techniques that might be constructed for this reason. Recently, AI, machine and deep learning is used in many various fields like agricultures, education, health, finance and business [3]. The employee attrition prediction using artificial intelligence has gained the researchers attention. Furthermore, the huge amount of dataset concerning this topic lead to more research in this area [4]. This study concentrates on the employee attrition prediction using deep machine and deep learning models, where the IBM analytics dataset has been utilized for the training, validating and testing the proposed models. The dataset contains 35 variables (features) from 1,470 employees. The class attrition has two possible outcome (current employees or former employees). This dataset is not balanced; that means the class (attrition) has 238 positive samples (previous worker) and 1,233 negative records (present worker). Because the data is unbalanced, the process of prediction becomes very difficult job.

2. OBJECTIVES

The objectives of this study are:

- To utilize machine and deep learning techniques with pre-processing steps to get better accuracy of attrition of employee.
- To analyze the dataset to be able to identify the variables(features)
- To detect the principal features in the dataset.
- To get accurate effects, we verified our proposed models with un-balanced and balanced datasets.
- To test the proposed models to identify the best technique of attrition of employees.

3. PROBLEM STATEMENT

Recently there is an increase attention to HR, as worker skills and quality represent factors of growth and a tangible modest benefit for businesses. Once proved its mettle in marketing and sales, analytics and AI is likewise becoming central to employee-related choices within HR management.

Organizational development mostly depends on staff retention. Dropping employees normally influences the self-esteem of the company and contracting with new employees is extra costly than retaining present ones. Thus we are proposing deep and machine learning techniques for predicting employee attrition.

4. PREVIOUS STUDIES

Employee attrition issue was studied by researchers from various viewpoints. A few researchers made an analysis of behavior of employees to disclose the causes of taking the decision to leave or stay in the business [6, 7].

Some researchers employed ML techniques to forecast attrition of employees based on data that belong to the employees like in [8], they employed few ML methods: KNN, Random Forests (RF), and SVM using various parameters. The researchers utilized 3 forms of the dataset (un-balanced, under sampled, and over sampled). Even though their over-sampled data exhibited great accuracy, original dataset accuracy insufficient.

The same dataset was used by researchers in [8] to make a comparison among a few ML methods, like, naïve Bayes, decision tree, and KKN for the prediction of employee attrition. The researchers tested the methods utilizing cross validation with 10-fold and 65% of the dataset for training and 35% for testing. Their work accuracy is not that good in comparison with other studies due to not utilizing data pre-processing phase.

The researchers in [10] analyzed the causes that influence a worker to think of leaving the business, wherever several ML methods were implemented to select the top classification model for employee attrition. The methods embrace logistic regression, KNN, naïve Bayes, and SVM. Their work was tested using training and testing data. Yet, the testing score was superior to the training score that is not bad, however, there is room for improvements.

The researchers in [11] offered a 3 phase outline for predicting employee attrition. The 1st phase, utilized “max-out” technique for the selection of features to reduce the dataset. Logistic regression model was used, training and testing for the prediction, in 2nd phase. The validation model and assurance investigation was accomplished in the 3rd phase. Their accuracies were low, their system was very complex because of the pre and post processing.

There researchers in [12] employed random forest and classification trees for predicting worker attrition. They began with pre-processed the dataset by omitting the less influential variables by Pearson correlation. Nevertheless, the work demonstrated

minor enhancement in accuracy in relation with the other ML techniques.

There researchers in [13] employed tree techniques for the prediction of attrition of employees. The used techniques contain light gradient boosted and random forests that achieved the highest accuracy. Private Dataset was used that consisted of 5,550 samples.

Other works, such as in [14], employed private dataset also, this prevent us from making a comparing their work with the current study. The accuracy of the preceding studies can be made better to obtain better accuracy and sureness.

Our proposed study use ML, DL and pre-processing of the data to enhance the accuracy of prediction.

5. METHODOLOGY

Our proposed ML and DL models examines the dataset to discover the most important variables that enhance the accuracy and create projecting model in line with these steps:

- Collecting the dataset about the employees.
- Pre-processing the collected dataset: Dataset is cleaned and normalized.
- Dataset analysis: The most important features are identified.
- Dataset balancing: Since the attrition feature is not balanced, Smote Function is used to balance it.
- Creating the proposed models: The appropriate structure for the models are designated to enhance the efficiency of the employee attrition prediction.
- Splitting the dataset: Split the dataset into(train, valid, and test)
- Train and validate the proposed model using the train and valid datasets
- Test the proposed models using the test dataset.
- Select the best model among the proposed models

5.1 Dataset Description

The dataset was collected from Kaggle Repository. It consisted of 35 variables from 1,470 workers. The variables of the dataset and variable types are shown in the Table 1. The output class is called (Attrition) that signifies the worker judgment whether leaving or staying in the company (True, False).

Table 1. Dataset features names and types.

Feature Name	Type	Feature Name	Type
Age	Numeric	Distance From Home	Numeric
Monthly Income	Numeric	Over Time	Categorical
Business Travel	Categorical	Education	Categorical
Monthly Rate	Numeric	Percent Salary Hike	Numeric
Daily Rate	Numeric	Education Field	Categorical
Number of Companies Worked	Numeric	Performance Rating	Numeric
Department	Categorical	Relationship Satisfaction	Categorical
Over 18	Categorical	Employee Number	Numeric
Employee Count	Numeric	Standard Hours	Numeric
Environment Satisfaction	Categorical	Stock Option Level	Categorical
Gender	Categorical	Training Times Last Year	Numeric
Total Working Years	Numeric	Job Involvement	Categorical
Hourly Rate	Numeric	Work Life Balance	Categorical
Job Level	Categorical	Years Since Last Promotion	Numeric
Years At Company	Numeric	Job Satisfaction	Categorical
Job Role	Categorical	Years With Current Manager	Numeric
Years In Current Role	Numeric	Marital Status	Categorical
Attrition	Categorical		

5.2 Pre-processing

Pre-processing operation is a vital phase in machine and deep learning that considerably increase the performance a model. Pre-processing comprises normalizing, cleaning and categorical data encoding that is going to be discussed in the coming sections.

5.2.1 Dataset Cleaning

Naive analysis of the dataset discloses that a few features are same for all workers like Employee-Count, Over-18, and Standard-Hours, thus we dropped them they have been omitted at this phase. Moreover, Employee-Number was dropped because it has values that are un-related to our current prediction.

5.2.2 Encoding of Categorical Features

The dataset has some features that of categorical type (group of values) instead of numeric. Categorical features in most ML techniques can't be utilized right away. The original data encompasses numerous features of type categorical like (Business Travel, Departments, Education Field, Gender, Job Role, Marital Status, and Over time). This type of feature should be mapped to numeric values. "One hot encoding" was employed to convert them, where a "one-hot binary vector" is allocated for every value. As an example, Business travel feature takes three values (rarely travel, frequently travel, and does not travel), it is mapped to (1, 0, 0), (0, 1, 0) and (0, 0, 1) respectively.

5.2.3 Dataset Standardization

Usually features vary significantly in relation to their values that suffers from low prediction accuracy as a feature having big values might have bigger weight when compared to the other ones. In machine and deep learning, models deals with small values much better than big values. In order to overcome this problem, rescale feature values are used. Among the most popular approaches of a feature re-scaling is called normalization. The values are re-scaled to be in a particular range. The current study, a feature values are re-scaled to be between [0, 1]. The normalization equation is shown in the following [15]-[18]:

$$X_i = (X_i - X_{min}) / (X_{max} - X_{min}) \quad (1)$$

5.3 Data Analysis

The analysis of dataset properties was done by connecting every feature to the class feature employee attrition.

From Figure 1, the employees with age under 22 are more likely to leave, followed by the age group between 22 and 29. After 30, employees' attrition rate goes down.

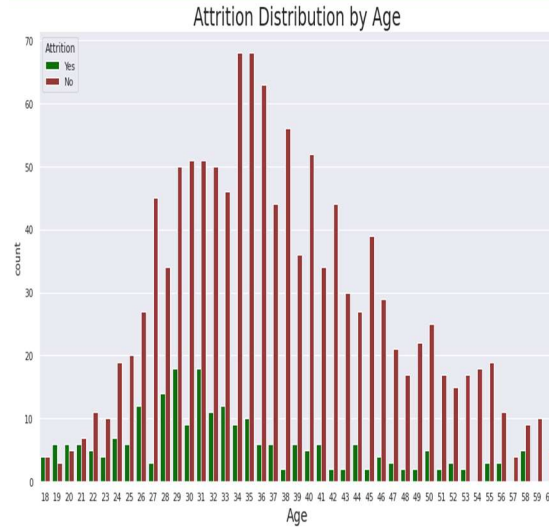


Figure 1: Attrition class distribution with relation to Age

Figure 2 shows that employees who travel frequently have a higher possibility to leave the company.

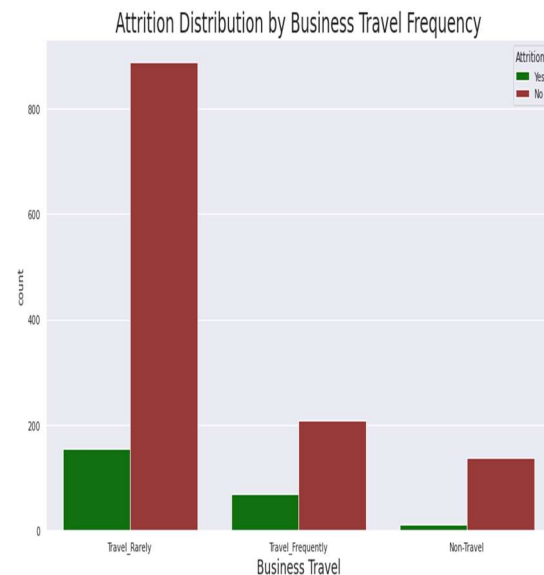


Figure 2: Attrition Distribution by Business Travel Frequency

Figure 3 shows no significant difference between attrition rates of different genders.

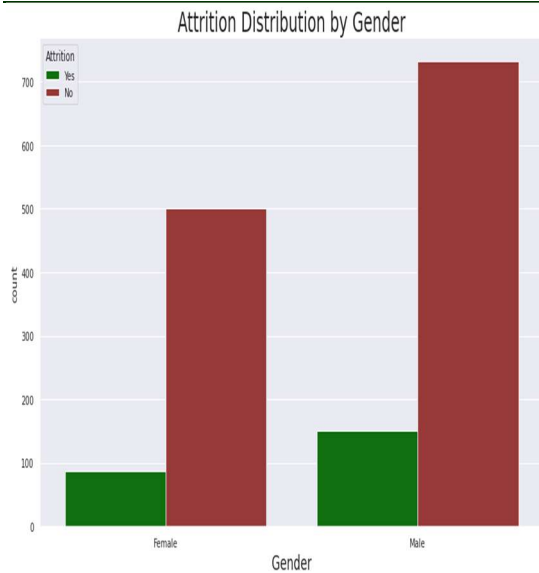


Figure 3: Attrition Distribution by Gender

The employees who have the lowest satisfaction with the working environment are more likely to leave the company as shown in Figure 4.

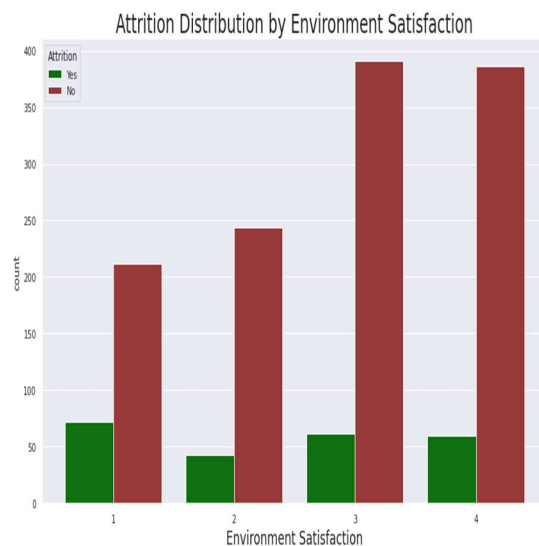


Figure 4: Attrition Distribution by environment satisfaction

Sales Department has the highest attrition rate as shown in Figure 5.

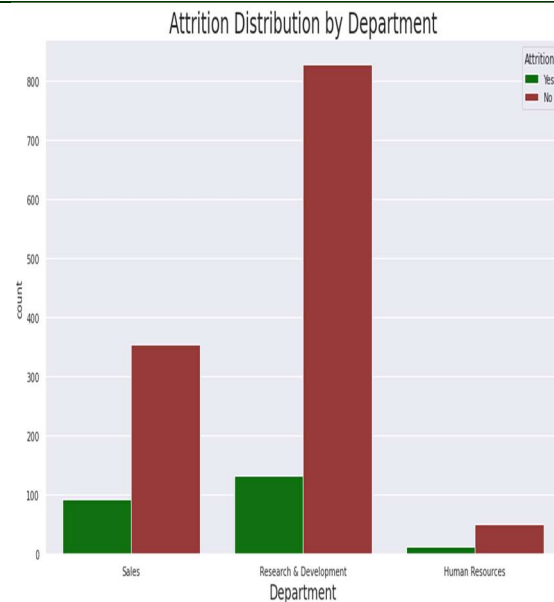


Figure 5: Attrition Distribution by Department

From Figure 6, the lower the job involvement, the higher the attrition rate.

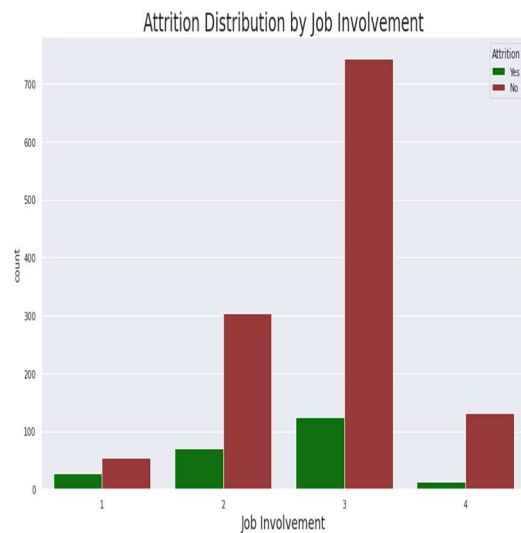


Figure 6: Attrition Distribution by job involvement

No significant difference shows between attrition rates of different performance ratings as can be seen in Figure 7.

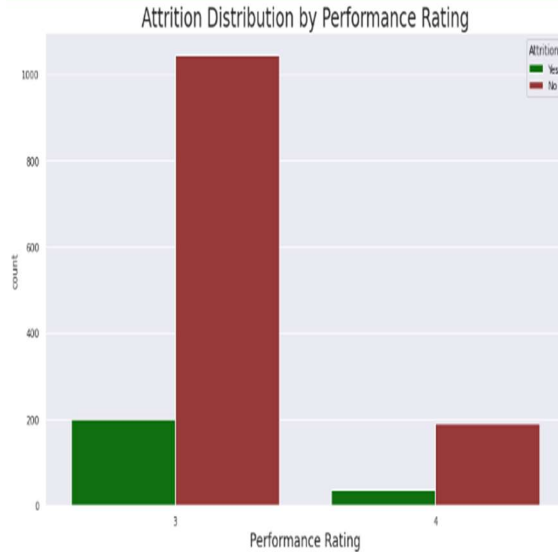


Figure 7: Attrition Distribution by performance rating

The employees without a stock option have the highest attrition rate, followed by those with the highest stock-option level as in Figure 8.

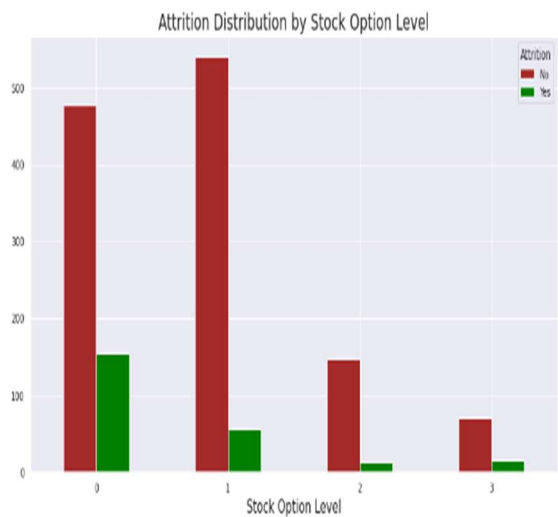


Figure 8: Attrition Distribution by stock option level

The employees who worked for less than two years have the highest attrition rate (almost 50%), followed by those who worked for 2 to 7 years as in Figure 9.

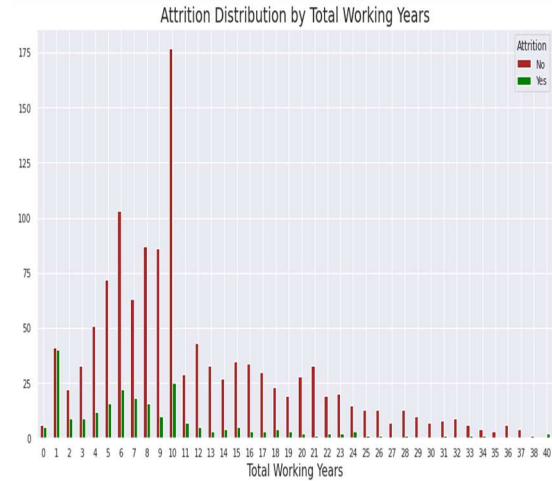


Figure 9: Attrition distribution by total working years

The less number of times training last year the higher attrition rate as illustrated in Figure 10.

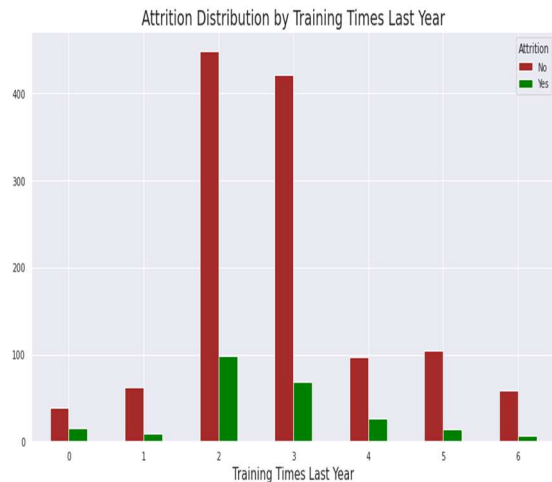


Figure 10: Attrition distribution by Training Times Last Year

Medium work-life balance brings down the attrition rate as illustrated in Figure 11.

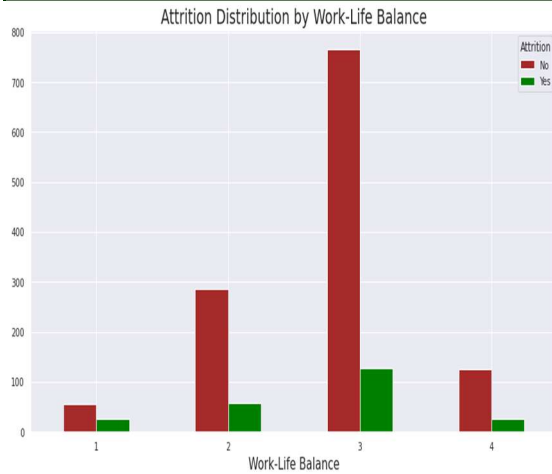


Figure 11: Attrition distribution by work-life balance

Sales Representatives have the highest attrition rate among all job roles in the company, followed by Laboratory Technicians as in Figure 12.

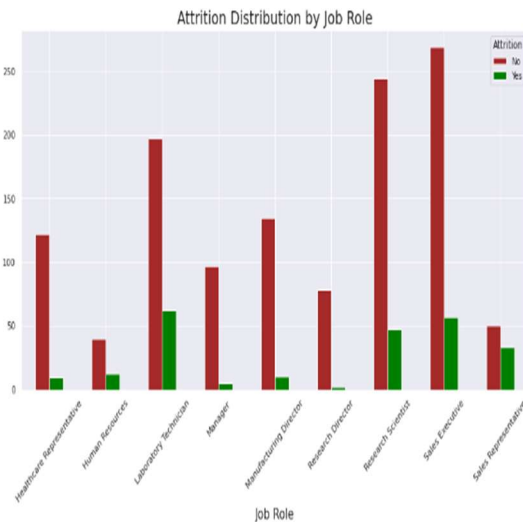


Figure 12: Attrition distribution by Job Role

The single employees are more likely to leave as in Figure 13.

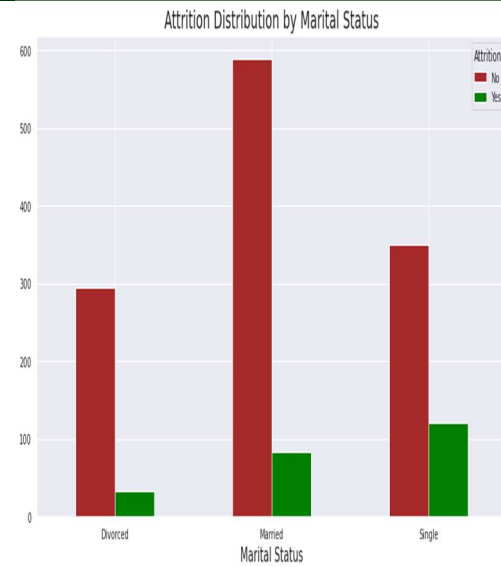


Figure 13: Attrition distribution by Marital Status

5.4 Models Building

The process of the modelling involve choosing methods that belongs to the different ML tools that we will use in the investigation. Nine ML models were selected: Random Forest Classifier, Extra Trees Regression, LGBM Classifier, Gradient Boosting Classifier, Label Propagation, Gradient Boosting Regression, MLP Classifier, Ada Boost Classifier and KNeighbors Classifier [19]-[22]. The aim was to recognize the top classifier for the analyzing the worker Attrition. Every model need consequently be trained and tested using the train and test datasets. The classifier with top accuracy results is then selected to be used for the worker attrition prediction. The models of ML techniques selected to be used in the current study are:

- Random Forest Classifier
- Extra Trees Regression
- LGBM Classifier
- Gradient Boosting Classifier
- Label Propagation
- Gradient Boosting Regression
- Ada Boost Classifier
- MLP Classifier
- KNeighbors Classifier

The result of testing the machine learning models are shown in table 2.

Table 2: result of testing the machine learning models

Machine Learning Model	Accuracies	Precisions	Recalls	F1_scores	Times-in Second
Random Forest Classifier	92.55%	92.56%	92.55%	92.55%	0.41
Extra Trees Regression	91.73%	93.42%	90.16%	91.76%	0.66
LGBM Classifier	91.25%	90.65%	92.38%	91.51%	0.21
Gradient Boosting Classifier	89.63%	90.35%	89.21%	89.78%	0.68
Label Propagation	88.33%	81.89%	99.05%	89.66%	0.22
Gradient Boosting Regression	89.30%	91.09%	87.62%	89.32%	0.60
Ada Boost Classifier	87.03%	86.15%	88.89%	87.50%	0.23
MLP Classifier	86.71%	85.63%	88.89%	87.23%	2.46
KNeighbors Classifier	84.77%	78.70%	96.19%	86.57%	0.06

Furthermore, we proposed one DL model for predicting worker attrition. This DL model consists of six layers: one for input layer, one for the output layer and four layers as hidden layers. The configuration of the proposed deep learning model is presented in Figure 14.

Layer (type)	Output Shape	Param #
Input_4(InputLayer)	(None, 31)	0
Dense_15 (Dense)	(None, 128)	4096
Dense_16 (Dense)	(None, 64)	8256
Dense_17 (Dense)	(None, 32)	2080
Dense_18 (Dense)	(None, 16)	528
Dense_19 (Dense)	(None, 2)	34
Total params: 14,994		
Trainable params: 14,994		
Non-trainable params: 0		

Figure 14: configuration of the proposed DL model

The training and validation accuracies and losses history of the proposed deep learning model is shown in Figure 15 and Figure 16.

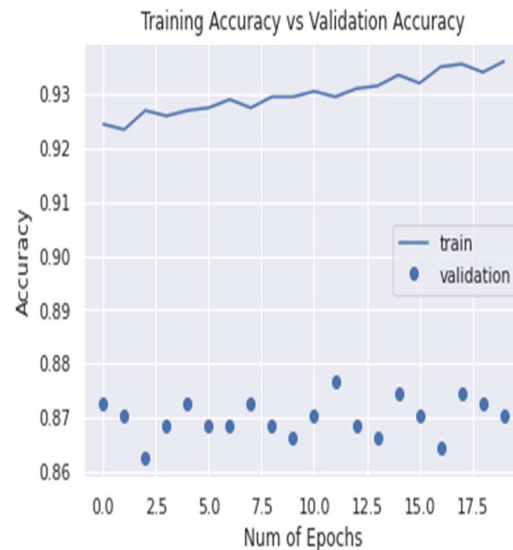


Figure 15 of the training and validation accuracies

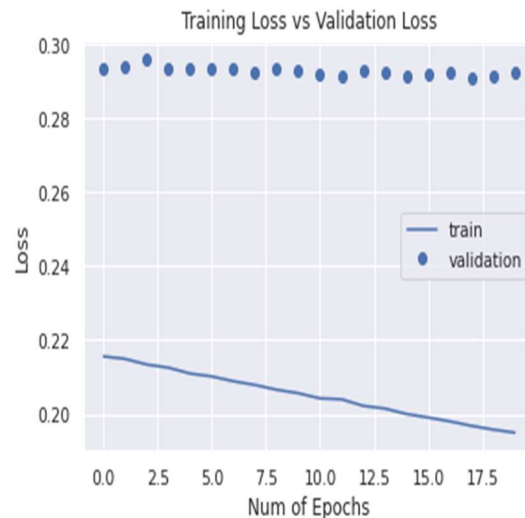


Figure 16: History of training and validation losses

The outcome of testing the proposed DL model is shown in Table 3.

Table 3: Outcome of testing the proposed DL model

Deep Learning Model	Accuracies	Precisions	Recalls	F1 scores	Time in second
Proposed DL model	94.52%	94.58%	94.52%	94.52%	0.39

5.5 Most Influential Features

After training, validating and testing all machine models, we picked the best model with highest F1-score, Random Forest Classifier, we got the most

important features using the function feature importances. From Figure 17, the most five top importance are Stock Option Level, Monthly Income, Job Satisfaction, Job Involvement, and Total Working Years.

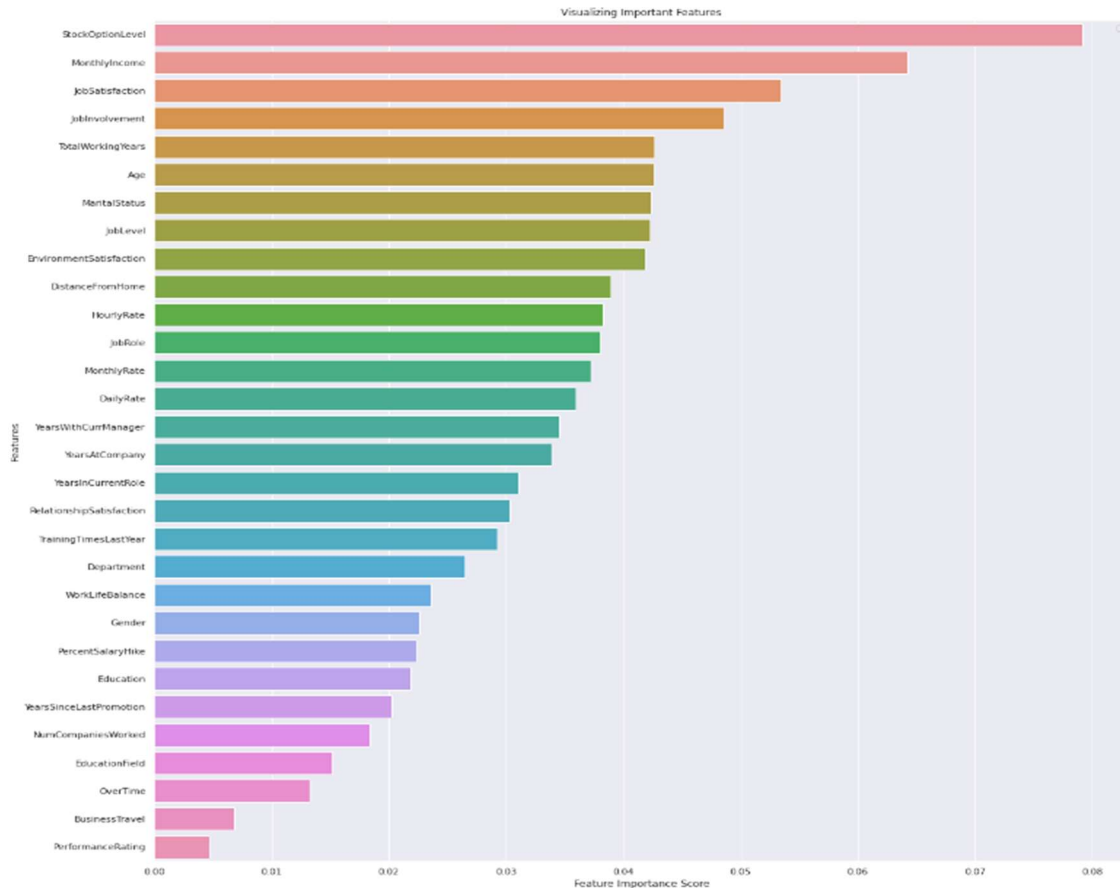


Figure 17: The most important features

6. LIMITATIONS AND FUTURE WORK

The current study was limited to 9 Machine learning models and one deep learning model. There are many machine models that could be testing in this study; however, we left it for futures studies.

7. CONCLUSIONS

The study aimed to help the department of human resources by giving them the required info about the probable decision for any employee that might be leaving the company. Our proposed model predicts if there is a latent risk of attrition by an employee.

The dataset was analyzed to get the best features that inspire the employee in leaving the company.

The top ML model was Random Forest Classifier: F1-score (92.55%), Precision (92.56%), Recall (92.55%) and time required for training, validating and testing the top ML model was (0.41 seconds).

On the other hand, the proposed DL model attained F1-score (94.52%), Precision (94.58%), Recall (94.52%) and time required for training, validating and testing the model was (0.39 seconds).

The proposed deep learning model has achieved better scores than the machine learning models used in this study and the study used in the previous studies [23]-[28]. After balancing the dataset, we trained, validated and testing all models. The proposed deep learning model has shown high prediction accuracy when compared to all other models. This due to the usage of deep learning techniques and due to appropriate utilization of pre-processing and choosing the most important features. The best top features are Stock Option Level, Monthly Income, Job Satisfaction, Job Involvement, and Total Working Years. The current study was able to answer all question raised in terms of implementing a model that detect the employee attrition with a high accuracy.

8. ACKNOWLEDGMENT

The authors are thankful to Palestine Technical University – Kadoorie for funding this research.

REFERENCES

- [1] Abunasser, B. S., et al., “Breast Cancer Detection and Classification using Deep Learning Xception Algorithm” International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 7, pp. 223-228, 2022. <http://dx.doi.org/10.14569/IJACSA.2022.0130729>
- [2] Obaid, T., et al. “Factors Contributing to an Effective E- Government Adoption in Palestine” Lecture Notes on Data Engineering and Communications Technologies, 127, pp. 663–676, 2022
- [3] Abunasser, B. S., et al., “Prediction of Instructor Performance using Machine and Deep Learning Techniques” International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 7, pp. 78-83, 2022. <http://dx.doi.org/10.14569/IJACSA.2022.0130711>
- [4] Saleh, A., et al. “Brain tumor classification using deep learning” Proceedings - 2020 International Conference on Assistive and Rehabilitation Technologies, iCareTech 2020, 2020, pp. 131–136, 9328072
- [5] Arqawi, S., et al. “Integration of the dimensions of computerized health information systems and their role in improving administrative performance in Al-Shifa medical complex” Journal of Theoretical and Applied Information Technology, vol. 98, no. 6, pp. 1087–1119, 2020.
- [6] Elzamly, A., et al. “Assessment risks for managing software planning processes in information technology systems” International Journal of Advanced Science and Technology, vol. 28, no. 1, pp. 327–338, 2019..
- [7] G. Gabrani, A. Kwatra. Machine Learning Based Predictive Model for Risk Assessment of Employee Attrition. In International Conference on Computational Science and Its Applications; Springer: Melbourne, Australia, pp. 189–201, 2018.
- [8] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, S. Hashemkhani Zolfani. An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. Mathematics, vol. 9, no. 1226, 2021.
- [9] M. Pratt, M. Boudhane, S. Cakula. Employee Attrition Estimation Using Random Forest Algorithm, Balt. J. Mod. Comput, vol. 9, pp. 49–66, 2021.
- [10] D. S. Rupa Chatterjee Das. Conceptualizing the Importance of HR Analytics in Attrition Reduction. Int. Res. J. Adv. Sci. Hub, vol. 2, pp. 40–48, 2020.
- [11] Saleh, et al. Brain Tumor Classification Using Deep Learning, 2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech). IEEE, 2020.
- [12] Setiawan, S. Suprihanto, A. Nugraha, J. Hutahaean, HR analytics: Employee attrition analysis using logistic regression. In IOP Conference Series: Materials Science and Engineering; IOP Publishing: Bandung, Indonesia, 2020; vol. 830, pp. 32-41, 2020.
- [13] S.Taylor, N. El-Rayes, M. Smith. An Explicative and Predictive Study of Employee Attrition Using Tree-based Models. In Proceedings of the 53rd Hawaii International Conference on System Sciences, Hawaii, HI, USA, pp.7–10 January 2020.
- [14] R. Yedida, R. Reddy, R. Vahi, R. Jana, A. GV, D. Kulkarni. Employee Attrition Prediction. arXiv 2018, arXiv:1806.10480.
- [15] Abu Naser, S.S. “Evaluating the effectiveness of the CPP-Tutor, an intelligent tutoring system for students learning to program in C++” Journal of Applied Sciences Research, vol. 5, no. 1, pp. 109-114, 2009.
- [16] Albatish, I.M., et al. Modeling and controlling smart traffic light system using a rule based system. Proceedings - 2019 International Conference on Promising Electronic Technologies, ICPET 2019, pp. 55–60, 2019, 8925318

- [17] Elzamly, A., et al. "A new conceptual framework modelling for cloud computing risk management in banking organizations" International Journal of Grid and Distributed Computing, vol. 9, no. 9, pp. 137–154, 2016.
- [18] Buhisi, N. I., et al. "Dynamic programming as a tool of decision supporting" Journal of Applied Sciences Research, vol. 5, no. 6, pp. 671-676, 2009.
- [19] Naser, S. S. A. "Developing an intelligent tutoring system for students learning to program in C++" Information Technology Journal, vol. 7, no. 7, pp. 1051-1060, 2008.
- [20] Fawzy A., et al. "Mechanical Reconfigurable Microstrip Antenna" International Journal of Microwave and Optical Technology, vol. 11, no. 3, pp.153-160, 2016.
- [21] Mady, S. A., et al. "Lean manufacturing dimensions and its relationship in promoting the improvement of production processes in industrial companies" International Journal on Emerging Technologies, vol. 11, no. 3, pp. 881–896, 2020.
- [22] Naser, S. S. A. "Developing visualization tool for teaching AI searching algorithms" Information Technology Journal, vol. 7, no. 2, pp. 350-355, 2008.
- [23] Fawzy A., et al. "Mechanical Reconfigurable Microstrip Antenna" International Journal of Microwave and Optical Technology, vol. 11, no. 3, pp.153-160, 2016.
- [24] Abu-Naser, S.S., et al. "An expert system for endocrine diagnosis and treatments using JESS" Journal of Artificial Intelligence, vol. 3, no. 4, pp. 239-251, 2010.
- [25] Kusuma S., et al. "BiDLNet: An Integrated Deep Learning Model for ECG-based Heart Disease Diagnosis" International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022.
- [26] Naser, S. S. A. "Intelligent tutoring system for teaching database to sophomore students in Gaza and its effect on their performance" Information Technology Journal, vol. 5, no. 5, pp. 916-922, 2006.
- [27] Shahd M., et al. "Prediction of Quality of Water According to a Random Forest Classifier" International Journal of Advanced Computer Science and Applications, vol. 3, no. 6, 2022.
- [28] Naser, S. S. A. "JEE-Tutor: An intelligent tutoring system for java expressions evaluation" Information Technology Journal, vol. 7, no. 3, pp. 528-532, 2008.