Productivity Forecasting of Employee Performance Using Machine Learning & Adaptive Neuro-Fuzzy Inference System

Abdalrhman M. Bormah[1*], Osman Taylan [2]

*Corresponding Author Email, aabormah@stu.kau.edu.sa

1,2-Department of Industrial Engineering, King AbdulAziz University, Jeddah, Saudi Arabia Kingdom.

Abstract
Employee productivity has been considered a crucial factor that threatens/diminishes company revenues and growth. Investigating the productivity of a garment company's workforce has been occupying this research's focus. Moreover, the garments industry is one of the most labor-intensive industries, and studying the actual productivity of employees is an important source for decision-making. Additionally, the productivity of the manpower is associated with influencers such as workload, incentives, overtime, and the capacity of the manpower versus task requirements. Based on theories and experiments, it has been found that employees' productivity could be affected by those mentioned factors and other variables such as the convenience of the surrounding environments, and workplace layout. Considering the power of artificial intelligence (AI) and machine learning (ML) techniques, starting by examining a set of regression algorithms, linear regression (LR), Regression Trees (RT), Support Vector Machines (SVM), Gaussian Processes Regression (GPR), and Ensemble of Trees Regression (ET) methods are used to predict the employees' productivity. Also, artificial neural networks (ANNs) are employed with a couple of training algorithms which are Levenberg-Marquardt (LM) & Bayesian Regularization (BR). The last application is the adaptive neuro-fuzzy inference system (ANFIS) via Hybrid and Backpropagation optimization. All the above models are studied to configure the impact of six independent predictors on productivity. In conclusion, medium regression trees give the RMSE of 0.10926 for training, and R-squared value of 0.69, Exponential Gaussian processes regression 0.10627 for RMSE for the training and 0.6 for R-squared respectively. The ANNs of Bayesian Regularization produced a value of 0.120476814 for RMSE and 0.72248 for R-Squared of the highest coefficient of determination.
.

Keywords- Employees' Productivity, Artificial Neural Networks (ANNs), Adaptive Neuro-Fuzzy Inference System (ANFIS), Regression Tree, Support Vector Machine, Gaussian Processes Regression, Employees' Performance.

INTRODUCTION

Imagine a company that would like to make significant revenues in a highly competitive market while they are failed frequently to meet the customers' demand as they have insufficient planning and poor forecasting of the needed workforce and associated productivity. As a result of such a weak planning strategy, the organization would lose some customers as they prefer to dedicate their loyalty to other companies capable of fulfilling their needs. Therefore, meeting customers' demand needs efficient forecasting of manpower, machines, materials, etc. In the following chapters, the manpower's productivity levels would be evaluated and forecasted through different algorithms aiming to reach accurate predictions of how the productivity of the employees can be improved and meet the needed demand to ensure customer satisfaction eventually. On the contrary, failing to meet targeted demand can impact customers' loyalty levels, as per customer surveys and feedback, most customers are disappointed with the company of the unavailability status of the products, feeling no more interest in the experience and purchasing more products from the same brand. Furthermore, consumers have included the waiting time to purchase or deliver the products can deteriorate their satisfaction.

Moreover, most of the companies within Saudi's market tend to use traditional methods and approaches of forecasting the workforce's productivity based on the current structure without considering implied factors. Noting that implementing a fair overtime pattern along with reducing the excess headcount, and assigning encouraging incentives for free-defect production can enhance the prediction outcomes. Also, ignoring paramount factors such as using smart solutions including AI approaches, trendy patterns, and seasonality movements could lead some of those companies and business firms to be exposed and affected by the consequence of inaccurate forecasting those mentioned previously. For that, planning the performance demand & requirements must be updated and supported by new research, studies, and approaches to be more efficient while the planners are forecasting the demand of the organization.

Therefore, one of the main reasons to run this study is to judge the productivity of the workforce fairly, optimizing the organization's expenses on hiring and ensuring a suitable headcount for each unit of the business. Why Artificial Intelligence? As our life has been inducted by the different use of artificial intelligence tools and applications like product recommendations, purchase predictions when we are shopping online, processes automation in the business, and more applications of artificial intelligence are expected to be implemented in the foreseeable future. Accordingly, the artificial intelligence approaches have proven their capabilities and accuracy as they facilitate the process, leading to accurate outcomes, and saving time and capital. Therefore, within this research a batch of artificial intelligence approaches within the machine learning algorithms based on a computational intelligence technique which are Regression Learner via the models of linear regression (LR), Regression Trees (RT), Support Vector Machines Regression (SVM), Gaussian Processes Regression (GPR), and Ensemble of Trees Regression (ET), while for Artificial Neural Network (ANN) part, two different training algorithms have been implemented those are Levenberg-Marquardt & Bayesian Regularization, considering the last application which is the Adaptive Neuro Fizzy Integrated System (ANFIS) as a part of control system design and analysis tools that is tested through both Hybrid and Backpropagation methods of optimization. All those models have been examined to provide the forecasting of the assigned company from the aspect of performance & productivity to identify the recommended measures related to the manpower capacity and specifying the employment and productivity demand of the following year besides implementing a deep comparison of the outcomes with explaining the strength and weakness of each approach versus the forecasted data by the traditional approaches to prove the effectiveness of each algorithm. After that, some recommendations and remarks will be included for possibly enhancing the planning of the workforce & productivity demand, as those might be useful for future research and studies. Beginning with Chapter I: an introduction of the topic including the background and purpose. Followed by Chapter II: an inclusive literature review, then Chapter III: Applications and Methodology. After that, Chapter IV: where the discussion and interpretation of the outcomes. Additionally, Chapter V: concludes the main findings and recommendations. Finally, Chapter VI: where the citation of references and sources.

LITERATURE REVIEW

A noteworthy that a conducive work environment provides comfort and safety to employees and that can be a motivating factor to improve their productivity. Consequently, a deficient work environment can be the main reason behind the negative productivity and performance levels[1]. Notably, the productivity definition is the efforts of an individual to convert the specified input(s) to a planned output(s) through identified methods and time limits. Also, productivity is depending on a group of factors including supportive policies, and compensations. Contrary, there are some obstacles like Incivility, and harassment, [2]. Financial benefits are essential as 65% of the employees' dissatisfaction is due to the low wages as per American Psychological Association [3]. Employee retention is achievable by keeping attention to the associated costs of training or

hiring to maintain the planned productivity levels [4]. A motivation mindset and culture must be implemented among organization departments and workforce as employee productivity is significantly related to how intrinsically motivated employees are to contribute to achieving the targeted figures and levels of their organizations [5].

Supervised Machine Learning is a sub-part of AI that is about applying an algorithm that is trained by using past data to predict future related figures [6]. The SVM has a set of advantages as it is processing the prediction of the data in a short time with high accuracy as it can train the linear and non-liners models. In contrast, the drawbacks of SVM include the long establishment of the model, and it is not highly recommended for large-scale data [10]. Additionally, the GPR algorithm employs the kernel method which provides the implied distributions of latent functions that define the relationships between inputs and outputs for the model. Also, it is performed based on the Bayesian perspective. As it optimizes the log margins of the dataset [11].

Highlighting that the traditional forecasting approach is using time series hypotactically to forecast future demand based on past demand. Consequently, the AI forecasting process considers exogenous factors [12]. For Obtaining excellent results via ANNs it is better to use non-linear models. Also, ANNs are described as black box that has many inputs that will be processed inside to produce single or multiple outputs based on intelligence computations algorithms. Moreover, the performance of ANNs depends on their parameters and architecture [13]. Although, computations of ANNs are unknown yet many researchers have developed various approaches with accurate outcomes, like Error back propagation (EBP) and Levenberg Marquardt (LM) algorithms [14]. ANFIS is where the fuzzy inference system (FIS) processes the data in the adaptive ANN framework. The FIS uses the if-then rules to convert the qualitative inputs through logical reasoning while adaptive ANN is working to design the patterns [15].

METHODOLOGY

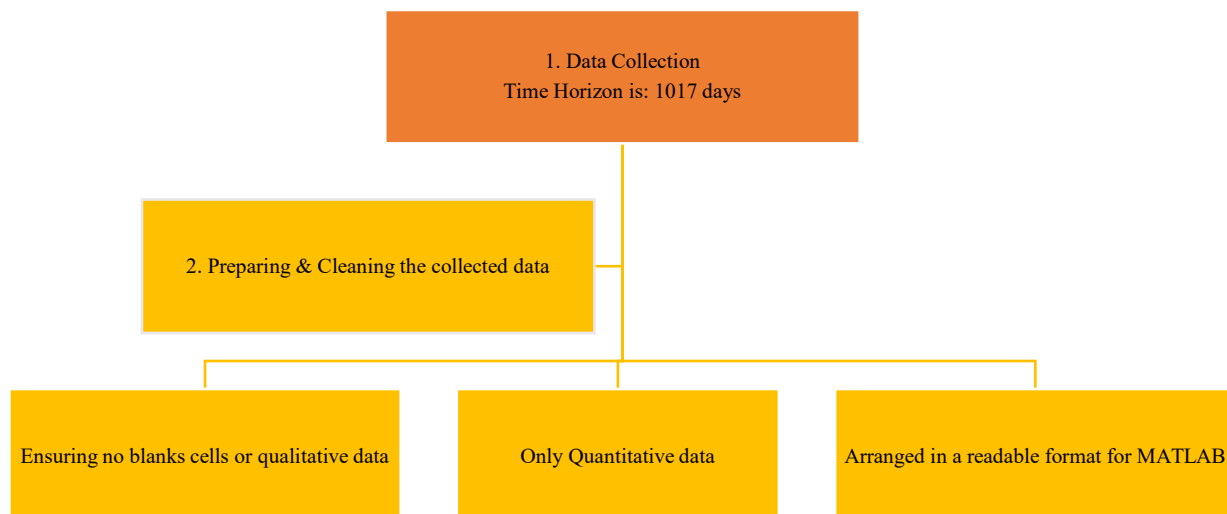*I.Data: collection, Preparing, and steps of implementing the training*

FIGURE 1
DATA COLLECTION & PREPARATION STEPS

targeted_productivity	SMV (Standard Minute Value)	WIP (Work in progress - Number of Items)	over_time	incentive	no_of_workers	actual_productivity
0.7	3.94	0	4320	0	18	0.759228395
0.8	4.08	0	1080	0	9	0.687555556
0.8	26.16	1073	10620	75	59	0.850502311
0.5	3.94	0	960	0	8	0.857916667
0.7	25.31	946	5760	60	54	0.700513566
0.8	15.26	961	4080	50	34	0.800401961
0.8	3.94	0	2400	0	10	0.753525
0.8	30.1	644	840	38	59	0.799982853
0.65	4.15	0	1440	0	8	0.821354167
0.7	11.41	685	3360	30	56	0.8211125
0.8	11.41	887	5490	50	30.5	0.800343766
0.8	2.9	0	960	0	8	0.667604167
0.6	15.26	1245	6360	23	34	0.60017284
0.75	4.15	0	1440	0	8	0.93532197
0.8	22.52	1500	6900	113	57.5	1.000018551
0.8	4.15	0	960	0	8	0.782447917
0.8	11.61	910	4470	38	31.5	0.80007657
0.8	4.08	0	1080	0	9	0.272
0.35	30.1	901	6960	0	58	0.350301724
0.8	11.61	955	6360	63	32	0.800191989
0.7	29.4	1146	6840	40	57	0.700614035
0.8	15.26	1133	4080	38	34	0.800401961
0.8	28.08	942	10530	75	58.5	0.850181818
0.8	11.61	632	0	50	57	0.800309211
0.8	26.16	1187	10620	75	59	0.850502311
0.75	18.79	1228	6240	45	52	0.750395513

FIGURE 2

DATA SAMPLE: INVESTIGATING EMPLOYEES' PRODUCTIVITY

II. Regression Learner (RL)

Cross Validation: Protect the model against the over-fitting, as it will use all the specified folds for training while leaving a single fold for the testing and it is recommended for small datasets.

Holdout Validation: by identifying a percentage of the training set to be used for validation and it is more efficient for large datasets.

Resubstitution Validation: no protections against the over-fitting status, as it uses the whole data set for training and testing at once. Therefore, the accuracy will be lower than previous alternatives.

After that, the model will be created as the response will be plotted and readings of the data will be scattered. then, specifying which model type will be used for the training as it can be a single model, a set of models within the same group, or all the regression models at once, those models are:

1) **Linear Regression (LR):** a statistical approach that is used for predictive analysis and defining the relationship between the predictors and responses of the model with the following mathematical formula:
$$Y = a + bX \tag{1}$$

2) **Regression Trees (RT):** an iterative approach of regression that divides the data into nodes, branches & leaves, and it is used for identifying the possible events and predicting the potential outcomes, computing the sum of squared errors for tree T as per the following formula:
$$S = \sum_{c \ \in leaves \ (T)} \sum_{i \ \in c} (y_i - m_c)^2 \tag{2}$$
Where the prediction of leaf C is:
$$m_c = \frac{1}{n_c} \sum_{i \ \in c} yi \tag{3}$$

3) **Support Vector Machine (SVM):** a supervised machine learning approach that is used usually for classification purposes, SVM is working on classifying the data by defining a line that is called a hyperplane which separates the data in a wide range as possible, Moreover, the data categories will be scattered on both sides of the hyperplane and the closest dot of each category is called a support vector while the distance between the support vector and hyperplane is called margin. Furthermore, the goal of SVM is to maximize the margin where the associated hyperplane is considered as optimal. Lastly, the SVM algorithm works well with small datasets.

Hyperplane: an n-dimensional line that separates the classes of SVM into two different sections, it is a line that identifies the boundaries of SVM classes.

The Mathematical aspect of SVM for the linear regression model is represented by the dual formula:

$$L(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) x_i' x_j + \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} yi(\alpha_i^* - \alpha_i) \qquad (4)$$

Based on support vector (SV) the following function is used to forecast new values:

$$f(x) = \sum_{i=1}^{N} (\alpha_n + \alpha_n^*)(x_n' x) + b \qquad (5)$$

4) **Gaussian Processes Regression (GPR):** non-parametric Bayesian approach that suits a small-size dataset. GPR models can be associated with capturing the uncertainty of the predictions. Additionally, Bayesian's approach considers the probability distribution of all possible values not specified ones like some machine learning alternatives. Considering the below formula:

$$P(f|X) \sim N(f|\mu, K) \qquad (6)$$

$$f = \big(f(x_1), \dots, f(x_N)\big), \mu = \big(m(x_1), \dots, m(x_N)\big) \ and \ K_{ij} = k\big(x_i, x_j\big) \qquad (7)$$

Where m: the mean function, and k is: Kernel or covariance function.

5) **Ensemble of Trees (ET):** a supervised learning approach but unlike the regression tree, the ensemble model is combining several trees to enhance the process of predictions of the outputs rather than utilizing a single kind of tree. Furthermore, the ensemble of trees algorithm can be over-fitted easily as it is very sensitive to outliers. Consequently, the ensemble of trees usually leads to excellent prediction besides high robustness as it minimizes the spread of the predicted data.

Bagging Formula:

$$S_L(.) = \sum_{l=1}^{L} w_l(.) \qquad (8)$$

Where w_l's are the weak learners, and the above equation is for a simple average of regression.

Gradient Boosting Formula:

$$S_L(.) = \sum_{l=1}^{L} c_l \times w_l(.) \qquad (9)$$

Where c_l's are the coefficients and the w_l's are the weak learners.

- The selected model will be ready to be trained, and initiating the training is followed by results on the left side of the window as the lower Root Mean Square Error (RMSE) model would be highlighted.
- Additionally, a test option will be available to be used after the training by testing a selected model or testing all trained models. Also, the testing stage starts by importing the data which will be tested.
- Interpreting the data visually through plots that include Prediction vs. Actual figures and Residuals (Validation) plots for both training and testing datasets. While an extra plot is available in case the user has applied the optimizable training model from the model type dropdown list which can be used to optimize the mean squared error (MSE) for any selected models.
- Finally, exporting the plots or the model besides generating the functions into the workspace.

III. Artificial Neural Network (ANN)

- By inserting the command nntool, the window of Neural Network Manager will appear.
- From the import button, we can import the inputs & outputs (target) from the workspace.
- From New button, we can define a Neural Network by specifying its parameters as training adoption learning functions, number of layers & neurons, and performance measures like MSE.
- Two training algorithms have been used as the following:

1) **Levenberg-Marquardt:** a 2nd-degree training algorithm for feed-forward ANN which provides a numerical value to minimize the non-linearity of the model by reducing the value of the sum of squared errors or residuals the LM algorithm can adjust the associated speed of the training. Not only that but LM can enhance the training by the steepest descent and quasi-Newton methods, it is one of the fastest training functions, as it's recommended for small or medium sizes datasets.

Considering the mathematical aspect of LM training, the objective function of LM training can be represented as follows:

$$E\big(w(n)\big) = 0.5 \sum_{t=1}^{Q} \sum_{r=0}^{N_1} \varepsilon_r^{2(L)}(t) = 0.5 \sum_{t=1}^{Q} \sum_{r=1}^{N_1} (y_r^{(L)}(t) - d_r^{(L)}(t))^2 \qquad (10)$$

Where Q: is the number of samples

$\varepsilon_r{}^{(L)}$: non-linear neuron error and it's defined as:

$$\varepsilon_r{}^{(L)}(t) = \varepsilon_r{}^{(Lr)}(t) = y_r{}^{(L)}(t) - d_r{}^{(L)}(t) \tag{11}$$

$d_r{}^{(L)}(t)$: the r-th expected vector of the i-th teaching sample, and it's calculated as the following:

$$\Delta\left(w(n)\right) = -[\nabla^2 E\left(w(n)\right)]^{-1} \nabla E(w(n)) \tag{12}$$

∇E: the gradient vector.

$$\nabla E\left(w(n)\right) = J^T\left(w(n)\right)\varepsilon(w(n)) \tag{13}$$

$\nabla^2 E\left(w(n)\right)$: is the Hessian Matrix.

$$\nabla^2 E\left(w(n)\right) = J^T\left(w(n)\right)J\left(w(n)\right) + S(w)n)) \tag{14}$$

$J\left(w(n)\right)$: The Jacobian matrix.

$$J\left(w(n)\right) = \begin{pmatrix} \frac{\partial y1}{\partial x1} & \cdots & \frac{\partial y1}{\partial xn} \\ \vdots & & \vdots \\ \frac{\partial yn}{\partial x1} & \cdots & \frac{\partial yn}{\partial xn} \end{pmatrix} \tag{15}$$

The training through LM algorithm applies the presenting of the next sample of data as computing the outputs of the ANN, applying the backpropagation with updating the rows of the Jacobian matrix via the following equation:

$$\frac{\partial \varepsilon_r^{(L)}(t)}{w_{ij}^{(l)}} = \delta_i^{(lr)}(t)\, x_j^{(l)}(t) \tag{16}$$

The Training will continue presenting the next sample till all samples are presented. Following that by calculating the error criterion, computing of weighted updated vector $\Delta\left(w(n)\right)$ through equation No.10 as updating the respective weights. After that, the criterion error is computed, and it will be compared with the previous error. Considering the new error value, if it was smaller than the previous one, it's a successful training, otherwise the training will be considered as a failure. The training ends when the target error limit is reached, or the training objective is achieved.

2) **Bayesian Regularization:** it is used to minimize the linear combination of squared errors and weights. In addition, this technique gives better results in terms of regression. Considerably, Bayesian Regularization is not an easy technique to be implemented. Moreover, Bayesian Regularization aligned with Levenberg-Marquardt as the backpropagation approach of neural network Jacobian J_X performance as the following formulas:

$$J_x = J_x \times J_x \tag{17}$$
$$J_e = J_x \times E \tag{18}$$
$$dx = -(JJ + I * mu)\backslash J_e \tag{19}$$

Where $E = $ the errors or the residuals, and $I = $ the identity matrix

- Next is to train the ANN after choosing the characteristics as No. of Epoch & Max Fail range.
- A set of parameters that have been processed will have appeared after training implementation.
- Additionally, a set of figures of the training process in the shape of Performance, Training State & Regression which interprets the correlations of coefficient (R-squared) plots will be available.
- After that, the outputs & errors figures are added to the Neural Network manager from where we can export them into the workspace and copy them to an Excel file to run needed computations before conducting the comparison between the actual and predicted figures.
- Note: All the figures of ANN are explained in detail under the results and discussion section.

IV. Adaptive Neuro-Fuzzy Inference System (ANFIS):

- from the apps– under the control system design and analysis– Neuro-Fuzzy Designer application.
- The window of the Neuro-Fuzzy Designer requires the user to upload a dataset in the load data area.
- Next to that, the generating of Fuzzy Inference System (FIS) options where the user can process the generation via loading it from a file or workspace, grid partition, or sub-clustering methods.
- After that, training of FIS options should be identified by selecting the optimization method to be hybrid or backpropagation, error of tolerance, and the number of Epoch (iterations).

1) **Hybrid:** a learning algorithm that combines the least square estimator and the gradient descent method.
2) **Backpropagation:** is a systematic approach to learning as it changes the weights of the neural network as per the activation function. Also, it is a supervised machine learning algorithm. As the outputs have some differences, they will be backpropagated to apply a set of adjustments to the weights between the neurons which minimizes the error.
   - Then, the testing dataset will be available by plotting the output against training, testing, or checking data.
   - On the right side of the plot, the number of membership functions of ANFIS can be shown.
   - Above the plotting space, there are three dropdown menus, where the user can export, import, or print the model. Also, the user can modify FIS properties, membership functions, or rules of the model. (All the details of the edit menu options will be explained in the results & discussion section). Lastly, visualizing the surface or the rules of the model will be obtainable.
   - Finally, the user should export the file with (.fis*) either to the workspace or to a file to be able to upload it again in the Fuzzy-logic designer and manipulate the rules, and conditions as needed.
   - The last step of ANFIS is exporting the predicted results and running the needed statistical computations.
   - The followings are the structure and the mathematical aspect of the ANFIS algorithm:
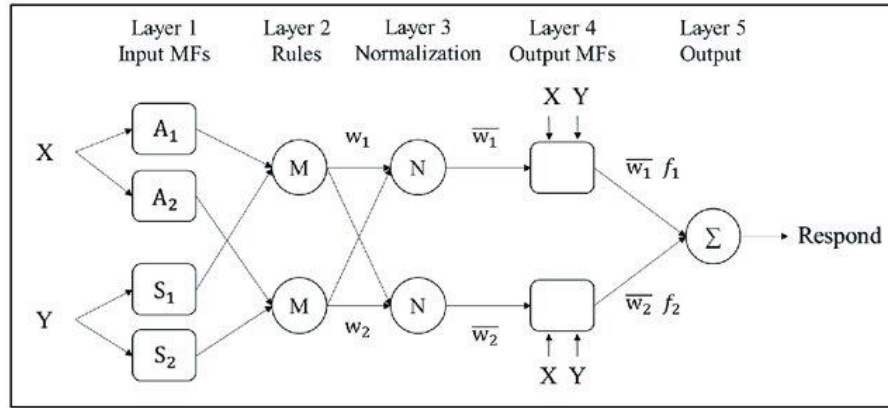   -

FIGURE 3
GENERAL STRUCTURE OF THE ANFIS FOR EMPLOYEES' PRODUCTIVITY INVESTIGATION (ALKHAZALEH, ET AL. 2022)

The ANFIS generally consists of 5 layers as the following:
- Layer 1: the input layer for the fuzzy system in which nodes represent a membership value to a linguistic term.
The Membership function types:

$$gaussian\ (c,\sigma) = e^{-0.5\left(\frac{x-c}{\sigma}\right)^2} \tag{20}$$

where c is the center of the function and σ represents the width.

$$Gaussian-bell\ shaped\ (a,b,c) = \frac{1}{1+|\frac{x-c}{b}|^{2a}} \tag{21}$$

where a is the width of the bell, b is a positive integer, and c is the center of the curve in a universe of discourse.

$$Triangular\ (a,b,c) = \max(\min\left(\frac{x-a}{b-a}, 1, \frac{c-x}{c-b}\right), 0) \tag{22}$$

where a & c are for the feet of the curve, while b is for the tip of the curve.

$$Trapaziodal\ (b,c,d) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \tag{23}$$

where a & b for the feet, and c & d for the shoulders of the curve.
- Layer 2: for the rules where each input node provides the strength of the rule by one of the multiplication operators AND or Prod to calculate and find the firing strength.

$$W_i = M\ A_i(x_0) * M\ B_i(x_1) \tag{24}$$

Where A_i is the linguistic value of the variable x_0, B_i is the linguistic variable of the variable x_1, and both terms are multiplied to find the firing strength of rule i.
- Layer 3: It is used for normalization purposes where the firing strength will be normalized by computing the ratio of the rule's firing strength to the sum of all rules firing strength through the equation:

$$\overline{W_i} = \frac{W_i}{\sum_{j=1}^{R} W_j} \tag{25}$$

where W_i is the firing strength of rule i[th] that is computed in layer no.2, and $\sum_{j=1}^{R} W_j$ is the sum of the firing strength for all rules.

- Layer 4: the adaptive node of the normalized firing strength of layer 3, can be computed via the following equation:

$$\overline{W_i f_i} = \overline{W_i}(p_i, q_i, r_i) \tag{26}$$

where W_i is the firing strength of rule i[th], and all p_i, q_i, r_i are the parameters of the node.

- Layer 5: it is a single node that represents the sum of all outputs from layer no.4, this node can be calculated through the below equation:

$$\sum_i \overline{W_i f_i} = \frac{\sum_i W_i f_i}{\sum_i W_i} \tag{27}$$

Where $W_i f_i$ is the output of layer no.4, and W_i is the firing strength of ith rule.

b. Analyzing the Data:

- Analyzing and comparing the different models of regression learner, ANN, and ANFIS.
- Comparing all the three applied methodologies, by considering the optimal model(s) to conclude which model and algorithm can be the best and why it has superior results among the other alternatives while considering the flaws of other approaches and why they should be avoided.
- The comparison will be based on the correlation of the coefficient R-squared values along with the accurate measurements of the values of root mean squared error (RMSE)

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{(n \sum X^2 - (\sum X)^2) * (\sum Y^2 - (\sum Y)^2)}} \tag{28}$$

Where r = Correlation of coefficient, n = the number of observations in the dataset, X= 1st variable Y= 2nd variable.
Note: in MATLAB, the considered coefficient value is R-squared which is the squared value of r.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \tag{29}$$

$$RMSE = \sqrt{MSE} \tag{30}$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)| \tag{31}$$

Where n = the number of observations in the dataset, Y_i = the actual or observed value.
$\hat{Y_i}$ = the predicted value, x_i = the data value in the set, $m(X)$ = the average or the mean of the dataset.

DISCUSSION AND RESULTS

As declared the algorithm of regression learner will be applied through different models to examine which might be the best according to the manpower productivity data. All the considered models have been trained by three different alternatives of cross-validation approaches of 3,5,10 folds, besides another three different hold-out percentages through the hold-out validation algorithm which are 10%,25%, and 40%. The comparison considers the following parameters: Root Mean Squared Error (RMSE), Correlation of Coefficient (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and training time in seconds. The examination has been implemented on six types of regression models those are Linear Regression (LR), Regression Trees (RT), Support Vector Machines (SVM), Gaussian Processes Regression (GPR), and Ensemble of Trees (ET), the following figures are summarizing each experiment and the associated results.

As seen below in Figure 4, we can conclude that the best models are the bagged Ensemble of Trees (BET) and Exponential Gaussian Processes Regression (EGPR) which they have the highest correlation factors of 0.47 and 0.45 respectively, besides recording the lowest values of RMSE of 0.12669 for ET, and 0.12956 for GPR model. Considering the time consumption, both ET and GPR are processing the model in a longer time within 3 folds cross-validation as the time in seconds is 2.9718 for ET and 2.4095 for GPR. Consequently, the worst results were for Linear Regression (LR) model as it has 0.24 for the correlation and 0.15259 as RMSE.
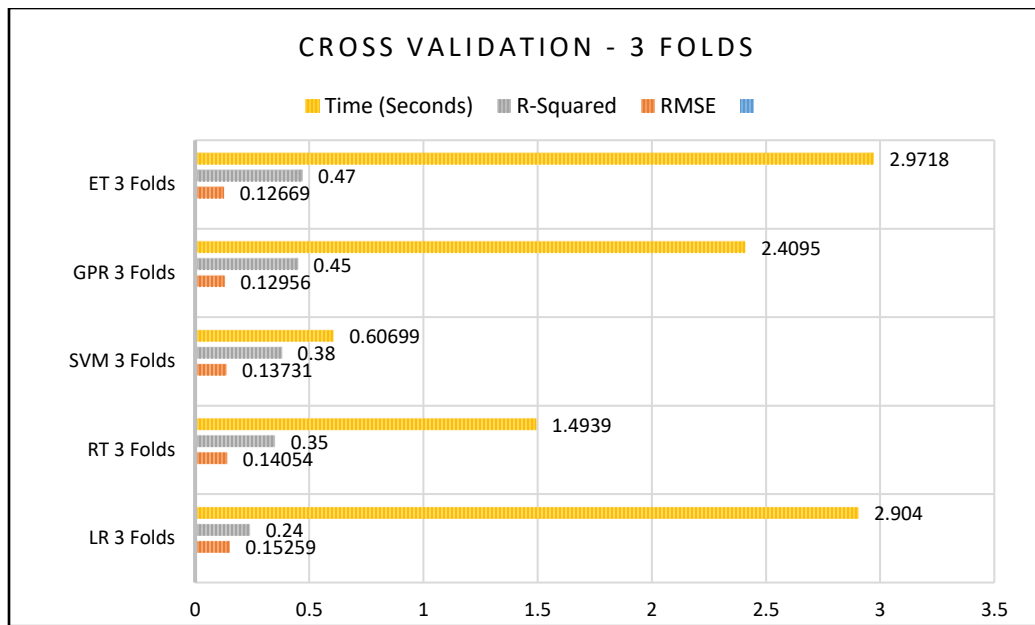
FIGURE 4:
EMPLOYEES' PRODUCTIVITY INVESTIGATION - 3 FOLDS CROSS VALIDATION TRAINING - REGRESSION MODELS
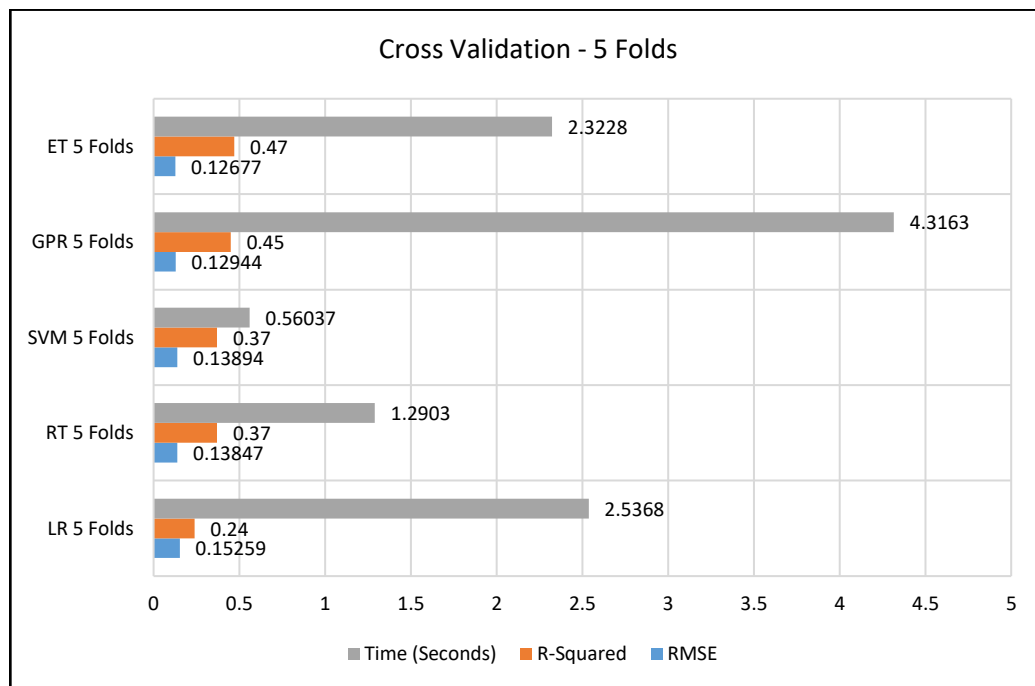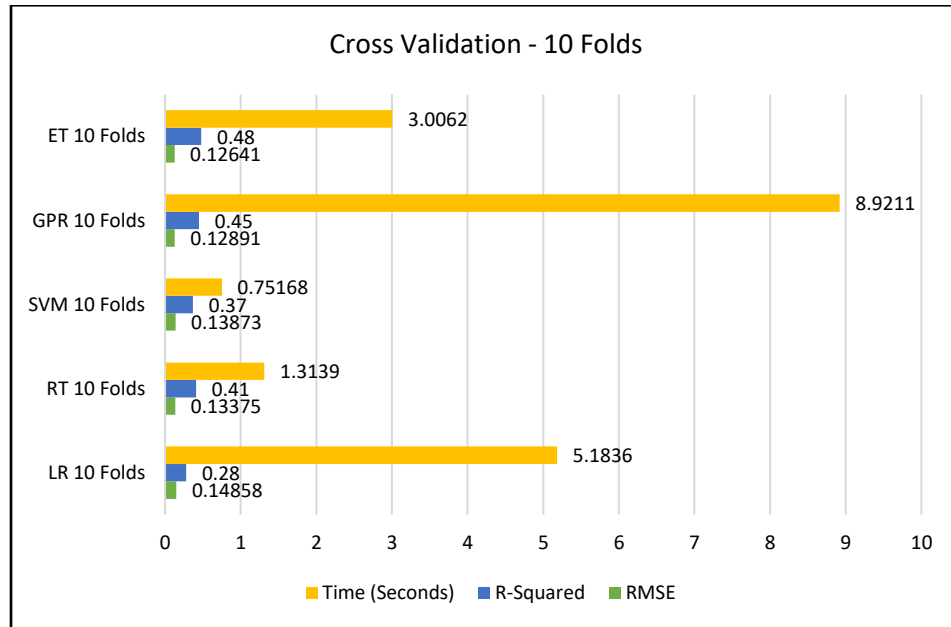

FIGURE 5
EMPLOYEES' PRODUCTIVITY INVESTIGATION - 5 FOLDS CROSS VALIDATION TRAINING - REGRESSION MODELS

As seen above in Figure 5, similarly as 3 folds validation, both BET and EGPR have recorded the best results as the associated RMSE values are 0.12677 and 0.12944 accordingly which are the minimum among all applied algorithms. Additionally, the R-squared value for BET is 0.47 and for EGPR is 0.45 those are the highest correlated percentages of the cross-validation with 5 folds. Noting that the consumed time for GPR is longer than BET as for EGPR the model has been

trained in 2.3163 seconds while it is 2.3228 seconds for ET. As the cross-validation with 3 folds, LR model has shown an unreliable correlation factor with a value of 0.24.

FIGURE 6
EMPLOYEES' PRODUCTIVITY INVESTIGATION - 10 FOLDS CROSS VALIDATION TRAINING - REGRESSION MODELS

As seen in Figure 6 above, noticing that BET and EGPR are the most appropriate models for the cross-validation approach in general as both have achieved the best outcomes of RMSE values of 0.0.12641 for BET and 0.12891 for EGPR as well as had the lowest RMSE even with 3- and 5-folds approach. In addition, both BET and EGPR like the 3-folds and 5-folds have recorded the highest correlation of 0.48 for BET and 0.45 for EGPR. On the other hand, we can conclude that LR is the weakest model when it comes to applying the cross-validation as it has the lowest correlation for the 10-folds the value of R-squared is 0.28.
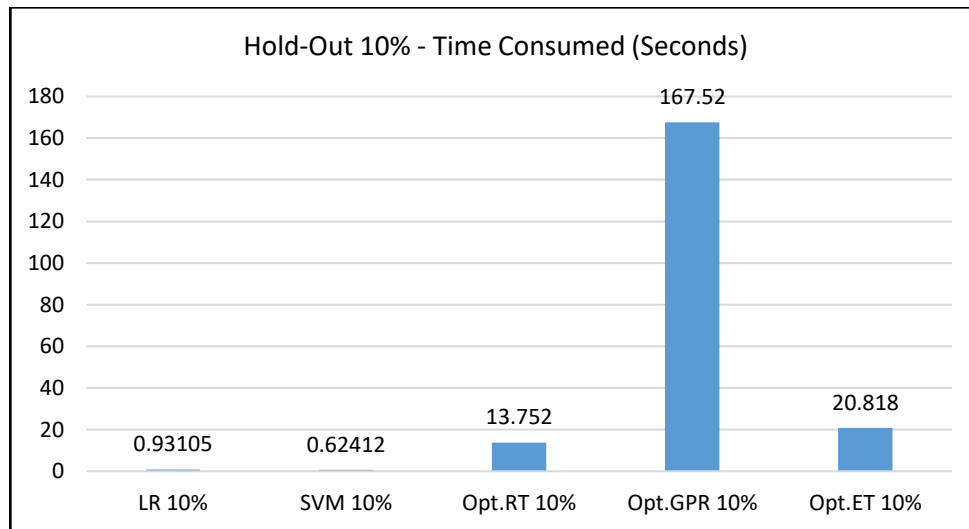

FIGURE 7
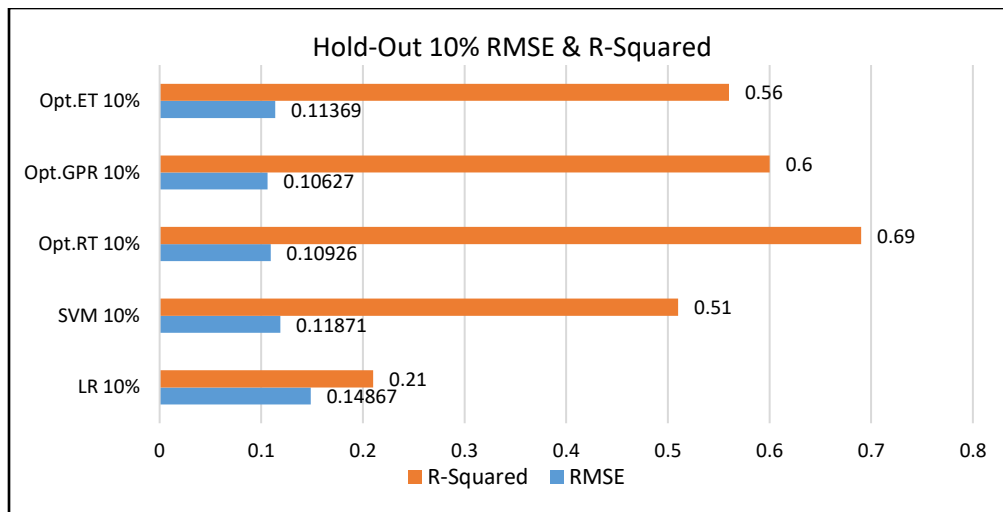EMPLOYEES' PRODUCTIVITY INVESTIGATION - 10% HOLD-OUT TRAINING - TIME CONSUMED (SECONDS)

FIGURE 8
EMPLOYEES' PRODUCTIVITY INVESTIGATION - 10% HOLD-OUT TRAINING - RMSE & R-SQUARED VALUES

As seen in both Figures 7 & 8, a noticeable improvement has been achieved generally from the aspect of RMSE and R-Squared values by applying the method of Hold-Out validation. Starting by Hold-Out with 10% of the dataset, the most recommended models are the optimized EGPR and Medium Regression Trees (MRT) as they have perfect results or R-Squared with values of 0.6 and 0.69 respectively beside very low RMSE with values of 0.10627 for EGPR and 0.10926 for MRT. Highlighting that optimized EGPR and MRT models have a significant variance regarding the consumed time for training as EGPR has trained the model in 167.52 seconds while the time for MRT is 20.818 seconds.
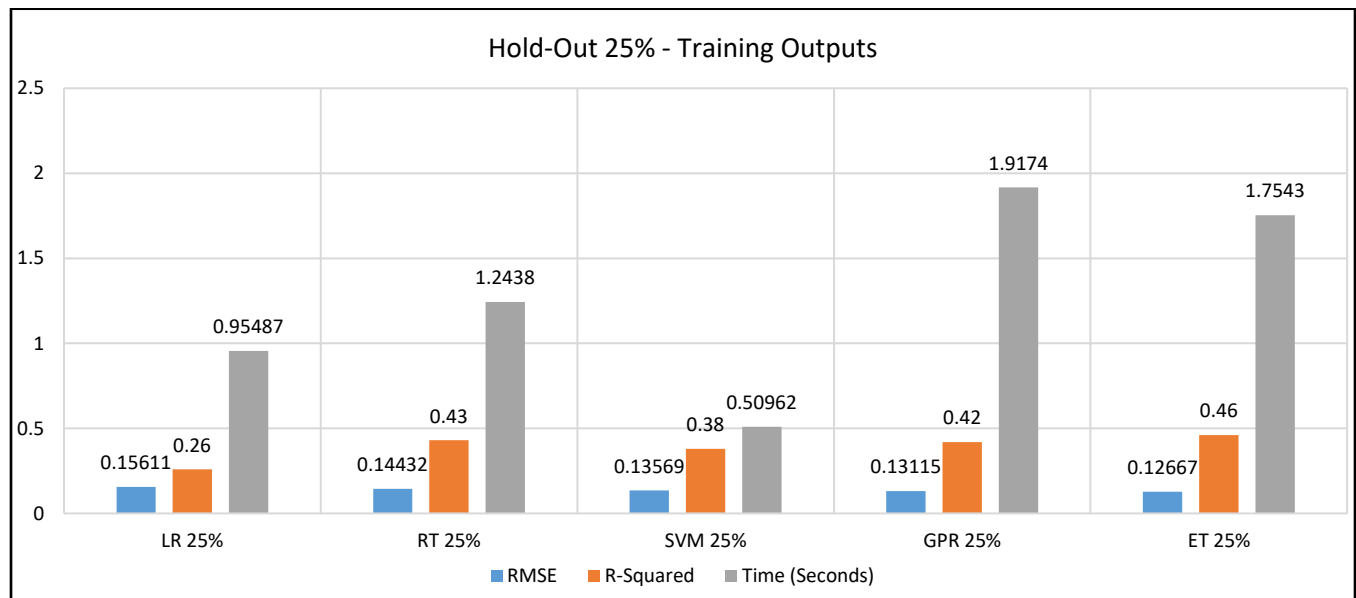

FIGURE 9
EMPLOYEES' PRODUCTIVITY INVESTIGATION - 25% HOLD-OUT TRAINING – RMSE, R-SQUARED & TIME CONSUMED (SECONDS)

As seen above in Figure 9, considering the second alternative of Hold-Out validation with 25%, the Bagged Ensemble of Trees (BET) model is the most suitable one to be implemented as it has defined the relationships between predictors and responses with a correlation factor with a value of 0.46 as it has the lowest residuals with RMSE value of 0.12667. Also, the model has been trained in a very short time of 1.7543 seconds through the BET model.
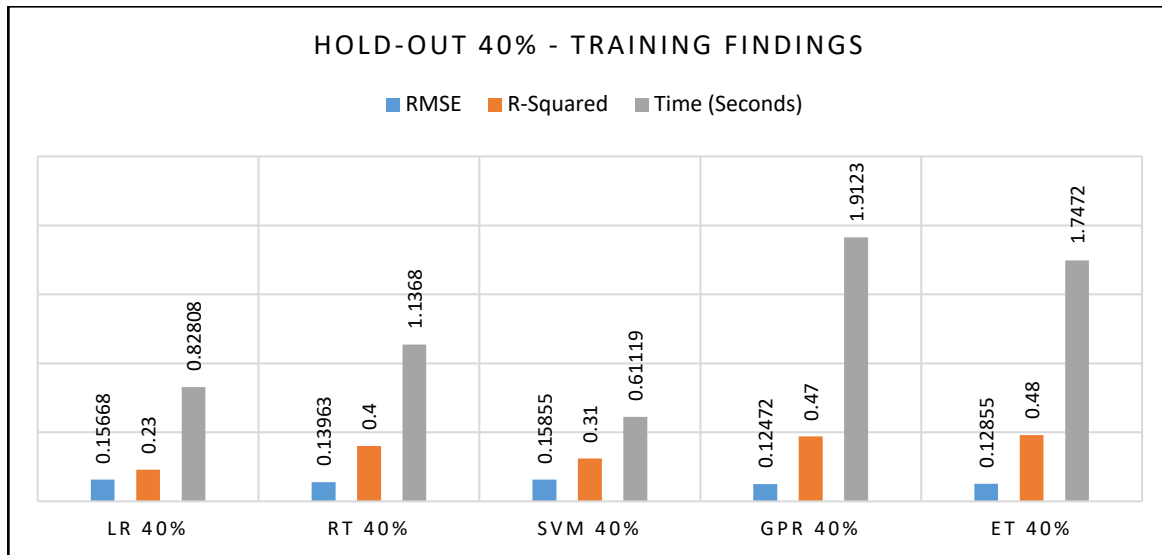
FIGURE 10
EMPLOYEES' PRODUCTIVITY INVESTIGATION - 40% HOLD-OUT TRAINING – RMSE, R-SQUARED & TIME CONSUMED (SECONDS)

As seen above in Figure 10, the last alternative of Hold-Out validation is with 40%, both the BET and EGPR are the most recommended models to be implemented as the correlation factor between inputs and outputs are 0.48 and 0.47 respectively. Mentioning that RMSE values are 0.12855 for BET and 0.12472 for EGPR which are the lowest. Both BET and EGPR have consumed a short time to conduct the training for the assigned model as the training take 1.7472 seconds for BET and 1.9123 for EGPR.
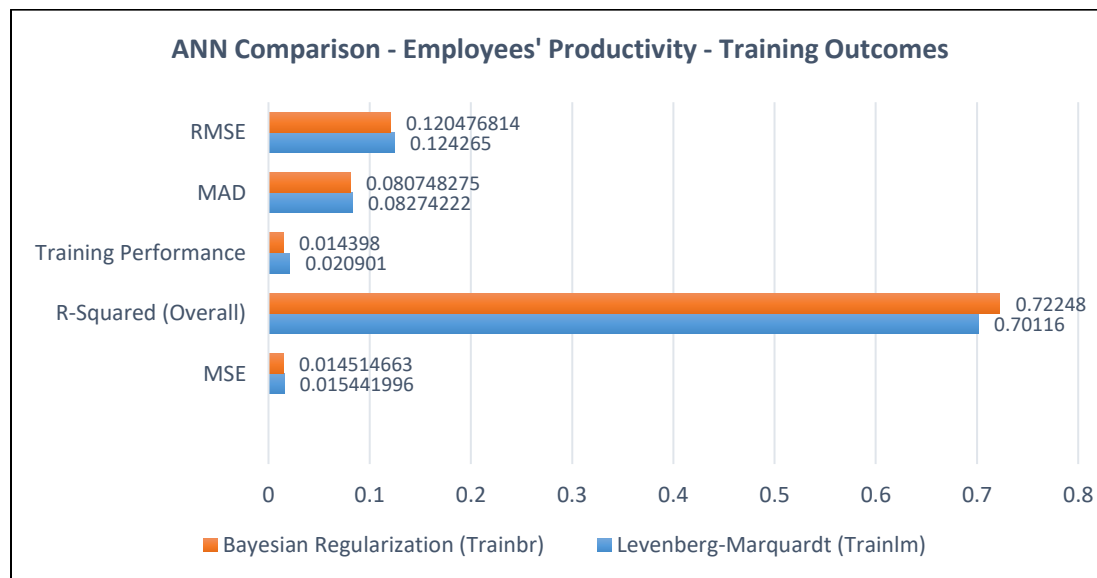

FIGURE 11
EMPLOYEES' PRODUCTIVITY INVESTIGATION – ANNs MODELS – TRAINING OUTCOMES

As seen above in Figure 11, a couple of training algorithms have been examined those are Bayesian Regularization (Trainbr) and Levenberg Marquardt (Trainlm), through using the feed-forward backpropagation method. Fortunately, both have perfect results as shown, starting with the RMSE values of 0.120476814 Trainbr and 0.124265 for Trainlm, while the mean absolute deviation (MAD) has a value of 0.080748275 for Trainbr and 0.08274222 for Trainlm. After that, the performance of

training has been measured as both Trainbr and Trainlm have achieved remarkable results of 0.014398, and 0.020901 respectively. Another important measurement has been applied to both algorithms that are the correlation factor R-squared, and both Trainbr and Trainlm have succussed in defining the relations between inputs and outputs with R-Squared values of 0.72248 for Trainbr and 0.70116 for Trainlm. The last statistical accuracy measurement that has been applied is the mean squared error (MSE), and the computed values are excellent as the errors of both Trainbr and Trainlm are 0.014514663 and 0.015441996 accordingly, which indicate how the variance between the actual and forecasted outputs are very small.

As seen below in Figure 12, the configuration of the ANFIS application of the assigned models, indicates the six predictors and a single response that is the actual productivity.
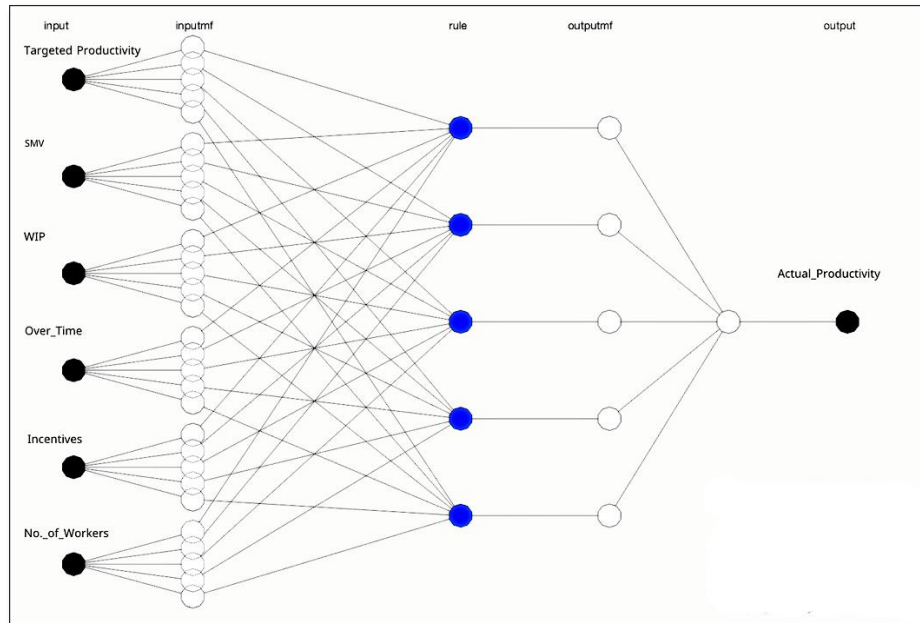

FIGURE 12
EMPLOYEES' PRODUCTIVITY INVESTIGATION - ANFIS CONFIGURATION

FIGURE 13
ANFIS TRAINING OUTCOMES – HYBRID AND BACKPROPAGATION OPTIMIZATIONS – RMSE VALUES

As seen above in Figure 13, the ANFIS algorithm has been investigated via a couple of optimization methods which are Hybrid and Backpropagation. To begin with, the Backpropagation shows an unacceptable level of errors as the best related RMSE value is 14.8752 and for that, it has been excluded from further statistical computations. As seen below in Figure 14, Considering the Hybrid approach, also the associated RMSE value is perfect with a value of 0.1205, yet it has failed in achieving a high correlation between the data as the R-Squared value is 0.2383 which is unreliable.
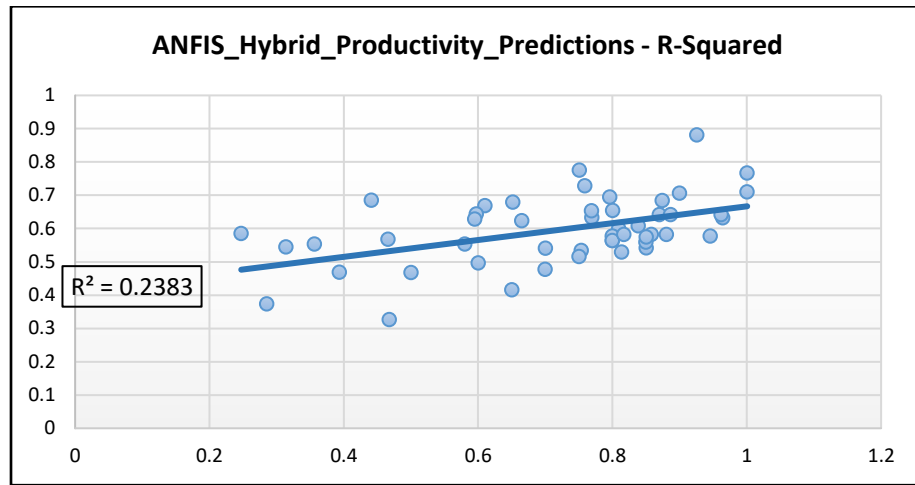
FIGURE 14
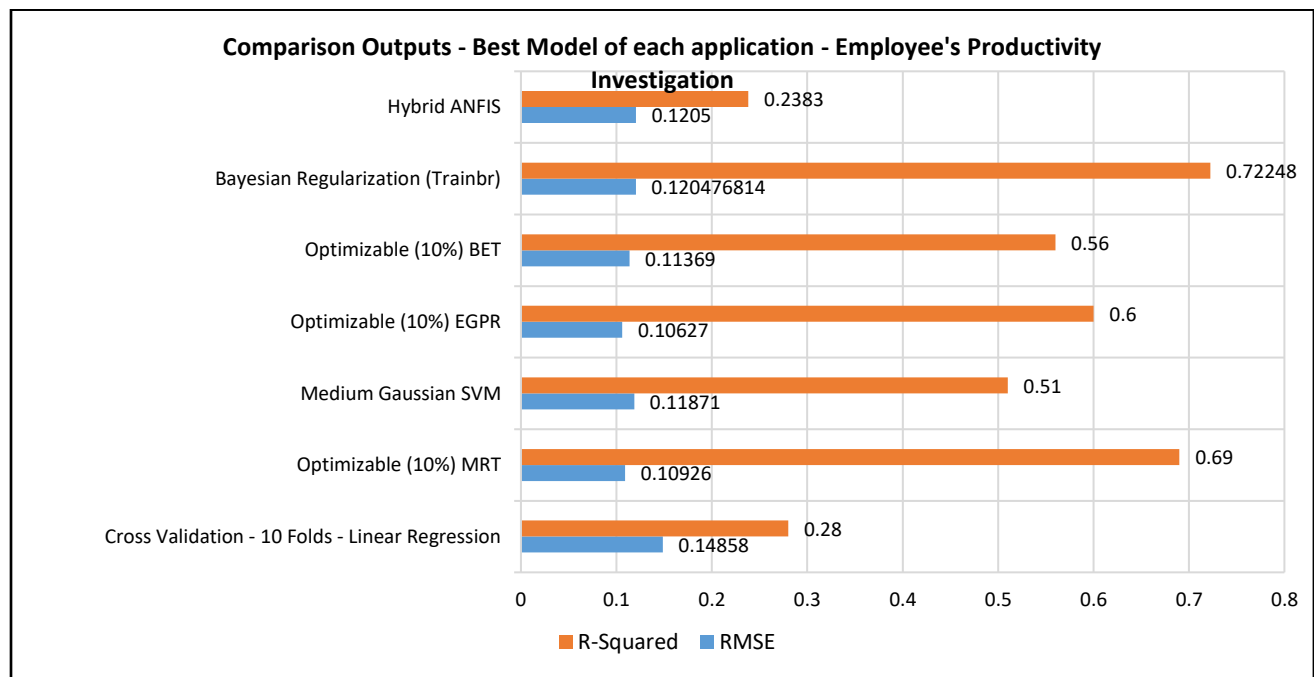ANFIS-HYBRID-PREDICTED EMPLOYEES' PRODUCTIVITY - R-SQUARED VALUE

FIGURE 15
EMPLOYEE'S PRODUCTIVITY INVESTIGATION - COMPARISON OUTPUTS - RMSE & R-SQUARED VALUES

As seen above in Figure 15, the comparison has been conducted including only the most recommended approach for each examination process. Additionally, the comparison is based on two statistical measurements those are R-squared and RMSE values. Starting with the ANFIS application, the Highest R-Squared and lowest RMSE values have been recorded by the Hybrid approach as they are 0.02382 and 0.1205 respectively. For ANN application the training through Trainbr shows perfect values of 0.72248 for R-Squared and 0.120476814 for RMSE. Next, is the linear regression application a couple of models are recommended which are the optimizable versions with 10% of Hold-Out validation for both EGPR and MRT as they have the highest correlation factors among the regression learner alternatives with values of 0.6 and 0.69 respectively. Also, both EGPR and MRT have recorded the lowest RMSE values of 0.10627 and 0.10926 accordingly. Based on low RMSE values and high correlation factors, it's highly recommended to apply either regression learner application via the medium

regression trees or Exponential Gaussian Processes Regression models, or the ANNs algorithm through the Bayesian Regularization (Trainbr). Highlighting the weakness of the ANFIS within the assigned dataset.

CONCLUSION

In conclusion, the productivity of employees has been examined through three different applications of artificial intelligence (AI) those are Regression Learner (RL) through five different approaches within cross-validation and holdout validation methods, Artificial Neural Networks (ANN) via a couple of training functions which are Levenberg-Marquardt (LM) and Bayesian Regularization (BR), and the Adaptive Neuro-Fuzzy Inference System (ANFIS) with two optimization approaches those are the hybrid and backpropagation. Moreover, the applied approaches are varied in the associated obtained outputs. In summary, the best models of each approach have been compared based on specific variables including RMSE and correlation coefficient (R-Squared). Firstly, the application of regression learner and among its models the optimizable held-out with 10% of both the medium regression trees (MRT) and exponential Gaussian processes regression (EGPR) are the best approaches compared to all implemented models as they have the lowest RMSE of 0.10926 and 0.10627 respectively, with 0.69 correlation for MRT and a value of 0.6 for EGPR. Secondly, the most accurate model of the neural network application is the training via Bayesian-Regularization function as it has lower computed values of RMSE = 0.120476814 along with a higher correlation coefficient of 0.72248 besides a better training performance of 0. 014398. Considering the last application, which is ANFIS as the hybrid training was better versus the backpropagation as it has RMSE value of 0.1205 yet it has an unreliable correlation coefficient with a value of 0.2383. Even though, all the mentioned RMSE values are perfect in predicting and forecasting the employees' productivity for the assigned dataset, the application of regression learner through the medium regression trees (MRT), exponential Gaussian processes regression (EGPR) through the optimizable held-out validation of 10% of the data, and the artificial neural network with Bayesian-Regularization algorithm of training (Trainbr) are the most recommended algorithms to investigate the assigned problems as they have the lowest RMSE values with the highest correlation between predictors and responses.

REFERENCES

[1] Basit, A. A., Hermina, T., & Al Kautsar, M. (2018). The influence of internal motivation and work environment on employee productivity. KnE Social Sciences.

[2] Anjum, A., Ming, X., Siddiqi, A. F., & Rasool, S. F. (2018). An empirical study analyzing job productivity in toxic workplace environments. International journal of environmental research and public health, 15(5), 1035.

[3] Mgbemena, G. C. Effects of Ergonomic Factors on Employees' Performance in the Brewery Industry: A Study of Nigeria Breweries Plc, Ama Enugu State, Nigeria.

[4] Mhatre, G., & Dhole, V. (2018). Trends in HRM: innovative technology for higher productivity of employees and the organizations. International Journal of Scientific and Engineering Research, 9(7), 1984-1990.

[5] Basahal, A., Jelli, A. A., Alsabban, A. S., Basahel, S., & Bajaba, S. (2022). Factors influencing employee productivity–A Saudi manager's perspective. International journal of business and management, 17(1), 39-51.

[6] Wenzel, H., Smit, D., & Sardesai, S. (2019). A literature review on machine learning in supply chain management. In Artificial Intelligence and Digital Transformation in Supply Chain Management: Innovative Approaches for Supply Chains. Proceedings of the Hamburg International Conference of Logistics (HICL), Vol. 27 (pp. 413-441). Berlin: epubli GmbH.

[7] Tohidi, H., Jabbari, M.M., (2012). "Measuring organizational learning capability". Procedia-social and behavioral sciences, 31, 428-432. https://doi.org/10.1016/j.sbspro.2011.12.079.

[8] Tohidi, H., Jabbari, M.M., (2012). "Important factors in determination of innovation type". Procedia Technology, 1, 570-573. https:// doi: 10.1016/j.protcy.2012.02.124

[9] Jabbari, M.M., Tohidi, H., (2012). "Providing a Framework for Measuring Innovation withinCompanies". Procedia Technology, 1, 583-585. https:// doi: 10.1016/j.protcy.2012.02.127

[10] Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification a review. Qubahan Academic Journal, 1(2), 81-90.

[11] Terry, N., & Choe, Y. (2021). Splitting Gaussian processes for computationally-efficient regression. Plos one, 16(8), e0256470.

[12] Cadavid, J. P. U., Lamouri, S., & Grabot, B. (2018, July). Trends in machine learning applied to demand & sales forecasting: A review. In International conference on information systems, logistics and supply chain.

[13] Al-Hasanat, A., Alasha'ary, H., Matrouk, K., Al-Qadi, Z., & Al-Shalabi, H. (2014). Experimental investigation of training algorithms used in back propagation artificial neural networks to apply curve fitting. European Journal of Scientific Research, 121(4), 328-335.

[14] Yu, H. (2011). Advanced learning algorithms of neural networks. Auburn University.

[15] Walia, N., Singh, H., & Sharma, A. (2015). ANFIS: Adaptive neuro-fuzzy inference system-a survey. International Journal of Computer Applications, 123(13).