Defence Science Journal, Vol. 74, No. 4, July 2024, pp. 496-504, DOI : 10.14429/dsj.74.19330

FuzzyBandit: An Autonomous Personalized Model Based on Contextual Multi-Arm Bandits Using Explainable AI

Nipun Bansal[#], Manju Bala[$] and Kapil Sharma[#,*]

[#] *Delhi Technological University, Delhi - 110 042, India*
[$] *I.P. College for Women, Delhi University, Delhi - 110 054, India*
[*] *E-mail: kapil@ieee.org*

ABSTRACT

In the era of artificial cognizance, context-aware decision-making problems have attracted significant attention. Contextual bandit addresses these problems by solving the exploration versus exploitation dilemma faced to provide customized solutions as per the user's liking. However, a high level of accountability is required, and there is a need to understand the underlying mechanism of the black box nature of the contextual bandit algorithms proposed in the literature. To overcome these shortcomings, an explainable AI (XAI) based FuzzyBandit model is proposed, which maximizes the cumulative reward by optimizing the decision at each trial based on the rewards received in previous observations and, at the same time, generates explanations for the decision made. The proposed model uses an adaptive neuro-fuzzy inference system (ANFIS) to address the vague nature of arm selection in contextual bandits and uses a feedback mechanism to adjust its parameters based on the relevance and diversity of the features to maximize reward generation. The FuzzyBandit model has also been empirically compared with the existing seven most popular art of literature models on four benchmark datasets over nine criteria, namely recall, specificity, precision, prevalence, F1 score, Matthews Correlation Coefficient (MCC), Fowlkes–Mallows index (FM), Critical Success Index (CSI) and accuracy.

Keywords: Explainable AI(XAI); Adaptive neuro-fuzzy inference system; Contextual multi-arm bandit; Personalization; Reinforcement learning

1. INTRODUCTION

In today's digital era, where a user is overwhelmed with information, presenting the user with the most relevant information, service, or product tailored as per the user's liking at the right moment and in a specific context has become the utmost priority for online applications. This personalized service or information improves the user's overall experience and engagement with the applications. With the advent of machine learning algorithms, a lot of solutions have been proposed to recommend items or services, but not much progress has been made when it comes to personalization. Contextual bandit algorithms address this issue and provide a way to optimise decisions by observing the payoffs associated with past decisions. With contextual bandits, it is possible to find a good trade-off between exploration(choosing the new option expecting a better payoff) and exploitation(choosing the known option) to get personalized decisions for every situation. In a nutshell, contextual bandits learn from user feedback to provide the best possible option on the fly, thereby reducing the risk of showing poor options to the user for too long by exploring the existing options and exploiting the most promising ones. Thus, contextual bandits are used successfully in real-life decision-making problems like clinical trials[12], recommender systems[13,37-40], dialogue systems[6], cognitive radio[32-33], brain and behavioural modelling[14], online advertisement[24], and many more.

Although these models can optimize decisions and provide accurate predictions, there is little explanation as to why and how a decision is made. Unfortunately, the black-box nature of the contextual bandits' algorithms is still unresolved, and many machine decisions are still poorly understood. Thus, there is a need for explainable AI-based models that interpret information in real-time settings, provide personalized solutions, and improve human understanding of the decisions made by the model. Further, the model should learn by itself without any human intervention so that it can be deployed at scale. Another major challenge faced is the cold-start problem[41], where many newly added items have very little past interactions with the users, thereby making it impossible to provide a personalized recommendation.

To meet these requirements, an explainable AI-based FuzzyBandit model has been proposed in this paper, where each arm in FuzzyBandit mimics an Adaptive Neuro-Fuzzy Inference System (ANFIS) to provide personalized choices to the user for the applications. The FuzzyBandit model is interpretable, i.e., the user can easily understand the underlying mechanism and justify the decision. It also allows the user to successfully answer questions such as why a particular action (or decision) is chosen at any instance of time or explain why things go wrong? Thereby, the user can easily understand the

Received : 01 August 2023, Revised : 05 October 2023
Accepted : 17 November 2023, Online published : 24 July 2024

decisions made by the proposed model. Further, coherent mathematics is also provided to calculate a trust score for the proposed FuzzyBandit model, which induces trust in the user regarding the decisions taken and detects erroneous reasoning in the proposed model, if any.

The paper is organised as follows: Section 2 reveals the existing literature on different types of contextual bandit models, their limitations, and their applications. An explainable AI-based FuzzyBandit model is proposed in section 3, and a detailed analysis of the benchmark datasets is described in section 4. The simulations and result analysis are carried out in section 5. Finally, the conclusions are given in section 6.

2. LITERATURE REVIEW

The multi-armed bandit problem[26] was first introduced as a sequential decision problem with statistical assumptions over the distribution of rewards over each arm to establish worst-case lower bounds for bandit experiments. It is recognized as an exploration/exploitation trade-off problem, maximizing the user's satisfaction by selecting the best arm, i.e. exploitation, while exploring the new choices/arms for uncertainties in the user's interests.Epsilon-greedy[21] and Epoch Greedy[18] are classic algorithms for random exploration, whereas Thompson Sampling[1,7] is the heuristic algorithm that handles the exploration-exploitation trade-off by maintaining probability distributions for each arm and then sampling from them on every trial to choose the one that predicts better rewards. LinUCB[8,20] assumed a linear relationship between the expected reward and the context. However, Chapelle & Li[7] showed that Thompson Sampling (TS) beats UCB.

With time, variants of the initial problem with different practical scenarios and constraints, such as non-stationary data (both data distributions and rewards may change with time), personalization on a per-user basis, and no assumption on how the rewards get generated, were introduced. In adversarial bandits[4], an adversary controls the rewards, and in stochastic bandit formulation[17], arms reward distribution is given by a well-behaved stochastic process instead of the statistical assumption as in the originally defined multi-arm bandit problem. Various optimal solutions using a stochastic formulation[17,29], adversarial formulation[3,30] and Bayesian formulation[5,16] have been provided in the literature.

Similarly, to provide personalized services, i.e. tailor the content as per the user needs and preferences, Contextual Multi-Armed Bandit Problem (CMAB)[10,18] uses a context feature vector (consists of the user's profile and choices) on each iteration and predicts the best choice out of the possible choices for user's satisfaction and interest. Recently, many solutions have been developed around contextual bandits to represent real-world problems. In the policy elimination algorithm[11], only good policies are kept in the working set, and the epoch-greedy algorithm is used for exploration. However, it is difficult to keep track of good policies and difficult to implement. If one removes an optimal policy by mistake, the algorithm can never recover. EXP-3[3] used weights for each arm, and weights were incremented exponentially for choosing the best arm. EXP-4[3] algorithm used advice from multiple experts to explore the connection between the context vector and the rewards of each arm, making it a suitable algorithm for non-stationary data.Banditron[15] model employed a perceptron[27]

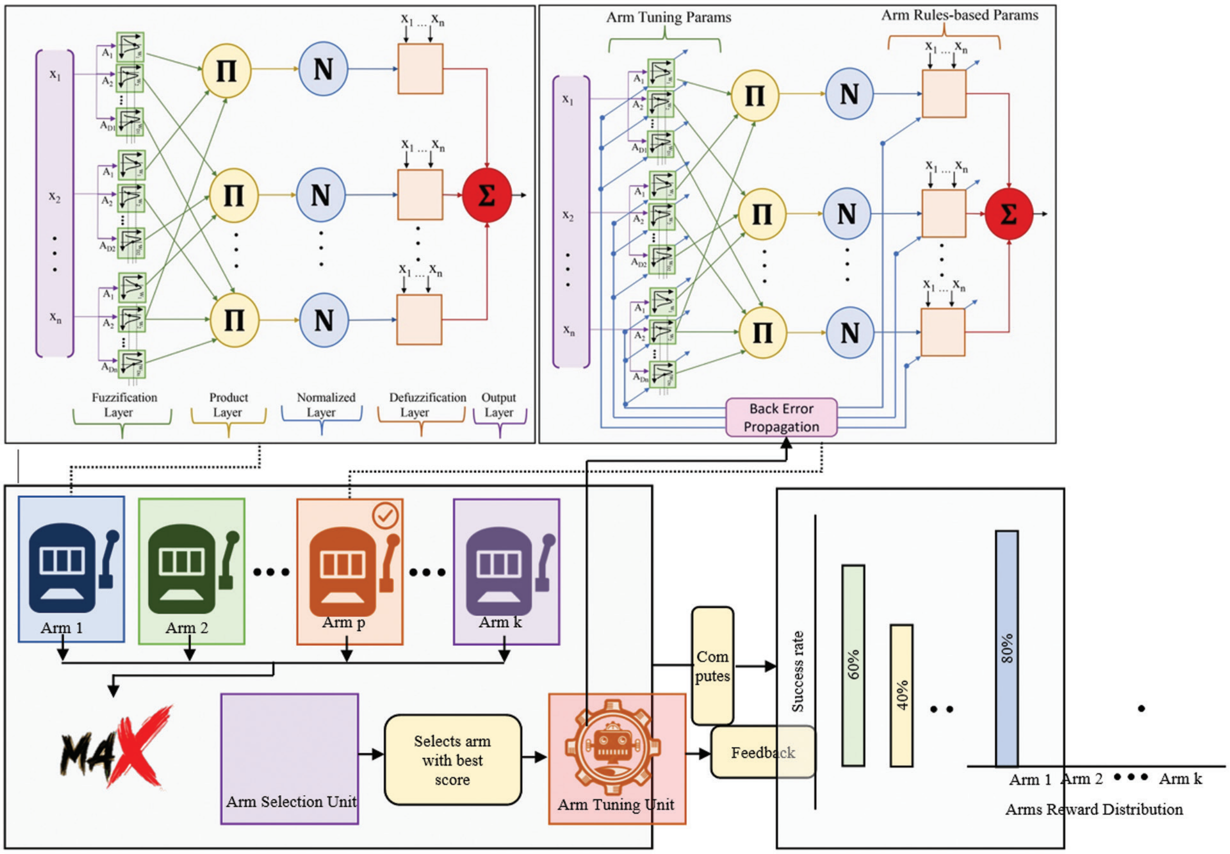

Figure 1. The proposed FuzzyBandit model.

to model expected rewards. It maintained weight vectors for each arm and outputs a prediction to the arm with the highest score. The algorithm employs a linear model. When banditron is coupled with upper confidence bound techniques, it is called Confidit[9].Confidit provides better performance than its base algorithm, banditron.

Linear algorithms lack representational power, and to overcome this shortcoming, deep neural networks became popular. Many deep neural networks use Thompson sampling as an exploration technique where a context is drawn at each round, and posterior distribution is updated with the result of the action, i.e. the feedback. An empirical comparative study[28] on how different posterior approximations by various algorithmic approaches/models, such as the Dropout model, neural linear model, etc., affect the decision-making performance via Thompson Sampling is presented. In deep neural network approaches, like Neural Bandits[2], a neural model is maintained for each arm for the ease of adding and removing arms. At each step, the context vector is taken as input for each neural model, and a score is obtained. The model chooses the arm with the highest score. Epsilon-Greedy is used as an exploration technique in this case. However, there are various limitations. Firstly, it is a very daunting task to train different architectures and to find optimal hyperparameters for each architecture, which requires a lot of computational power. Secondly, a neural network requires a large training dataset, which is infeasible in real-time applications where it is essential to respond quickly. The dropout[25] model addresses the above issues by dropping out the hidden layers in the network, i.e. randomly zeroes out the output of a neuron in the forward pass with a probability *p*. This prevents both overfitting and computational power to train the model. However, the number of hyperparameters to tune and the training time required will still be much higher when compared to the linear models.

3. METHODOLOGY

In this paper, it is proposed that a novel explainable AI (XAI) based FuzzyBandit model usage, an autonomous decision system that not only optimizes and personalizes decisions for every situation based on the previous observations but at the same time generates explanations for the decisions made. For every decision, the model generates a confidence score *a*, which enables the end user to easily understand and trust the decisions taken by the model. The proposed model simulates contextual multi-arm bandit (CMAB) model settings where it inputs a finite n-dimension user context feature vector, $x(t)=\{x_1(t), x_2(t),\ldots x_n(t)\}$ and chooses an action, $a(t)$ from an alternate number of choices/actions for each trial *t*. Each action is associated with a reward that is unknown to the model and is revealed after the action is chosen. The proposed model observes a binary reward, $r_{a(t)}(t)$ i.e. +1 if the rendered action is accepted by the user; otherwise, 0.For each trial, *t* let $a^*(t)$ be the optimal arm/choice which would yield the highest reward, $r_{a^*(t)}(t)$. Then the *regret*(*t*) is defined as the difference between the maximum reward for trial *t* and the observed reward by the model, i.e.

$$regret(t)= r_{a^*(t)}(t)- r_{a(t)}(t) \text{ I } x(t) \quad (1)$$

At the end of the *T* trials, the model aims to minimize the regret or maximize the total reward observed. This CMAB property allows the proposed FuzzyBandit model to simulate real-world problems for e.g., in an online food ordering app, each time the user logs into the app, the model uses the information about the user and renders a serviceable restaurant (i.e., action) to the user from a set of possible restaurants partner advertisement (modelled as *k* arms).The context feature is the information about the user: the user's past order history, favourite restaurants, average order values, a device used to order food, etc. The user feedback i.e. click/ no click on the advertisement, will act as a reward.

As depicted in Fig.1, the proposed model consists of *k* fuzzybandit arms where each fuzzybandit arm (FBA) corresponds to an action from an alternate number of choices/ actions possible. Each FBA represents a standalone ANFIS model developed using the Sugeno fuzzy model[19,23] and inputs an n-dimension user context feature $x(t)$. Since the inputted context feature vector is vague or imprecise, $x(t)$ is represented with more than one fuzzy set with membership functions, A_{11},A_{21},…A_{Dn}, to accommodate the possibility of more than one linguistic variable associated with the feature. For each membership function, F maps the element of $x(t)$ to a value between 0 and 1 i.e. F($x(t)$)→[0,1]. We have used the gaussian membership function for each node, as shown in the Fuzzification layer in Fig. 1 and is defined as:

$$O_{i,j}^{1} = e^{-\frac{\left(x_i-a_{i,j}\right)^2}{2b_{i,j}^2}} \qquad \forall\, i \in [1,n], j \in [1,\mathrm{m}] \quad (2)$$

where, $\{a_{i,j},b_{i,j}\}\epsilon RXR$ are arm-tuning parameters associated with each node and get updated with each iteration. *n* is the number of features of the user context feature vector, and *m* represents the number of membership function associated with each feature. The number of membership functions for each feature is data-dependent and is calculated experimentally. Next the real-time relevance scoring, $FBA_i(t)$ for each action associated with FBA is calculated using the FuzzyBandit Ranking Algorithm. The $FBA_i(t)$ score is calculated as:

$$FBA_i(t) = \sum_{j=1}^{m} W.\beta \quad (3)$$

where, β will be $\sum_{i=0}^{n}\left(q_{j,i} * x_i(t)\right)$, $q_{j,i}$ are arm-rules based parameters and *W* is the output from the previous layer for each FBA.

The arm-rule based parameters for each FBA is computed as the best solution for Eqn. (3) which is minimizing the square error $||Z \ - W\beta\ ||$. It is given as:

$$\beta^*=(W^TW)^{-1}W^TZ \quad (4)$$

where, W^T is the transpose of W and $(W^TW)^{-1}W^T$ is the pseudo inverse of W.

The $FBA_i(t)$ score calculated in Eqn. (4) is inputted into the FuzzyBandit Arm Selection Unit (FB-ASU).FB-ASU ranks each action, $a_i(t)$ corresponding to the arm, *i* based on the $FBA_i(t)$ score in the descending order i.e. the arm with the highest $FBA_i(t)$ the score gets the lowest rank. The lowest ranked arm is then outputted as the best possible arm for the given $x(t)$ by the proposed model and fed into the FuzzyBandit Arm Tuning Unit (FB-ATU). For each trial, FB-ATU checks

Algorithm 1 FuzzyBandit Ranking Algo

```
Input: Given data-set of the training sample $\{x(t),t\}_{i=1}^{n} \in \{R^d \times R\}$
Initialize $ANFIS_1(a_1,b_1)...ANFIS_k(a_k,b_k)$ for each arm        ▷ $a_k, b_k$ are arm-tuning parameters
for $t = 1,2,...$ T do
    for $arm, k = 1,2,\ldots,K$ do
      // FBA fuzzification
        for $i = 1,2,\ldots,n$ do
            for $j = 1,2,\ldots,m$ do
                $O^1[i][j] \leftarrow exp[-(x_i - a_{i,j})^2/2 * b_{i,j}^2]$

      // FBA production
        for $j = 1,2,\ldots,m$ do
            for $i = 1,2,\ldots,n-1$ do
                $O^2[j] \leftarrow product(O^1[i][j], O^1[i+1][j])$

      // FBA normalization
        for $j = 1,2,\ldots,m$ do
            $O^3[j] \leftarrow divide(O^2[j], totalsum(O^2))$

      // FBA defuzzification
        for $j = 1,2,\ldots,m$ do
            $O^4[j] \leftarrow product(O^3[j], F(x_1,x_2,..x_n,q_0,q_1,..q_n)$        ▷ $q_0,q_1,..q_n$ are arm-rule based
parameters and F is linear function
            $FBA\_score[k] \leftarrow FBA\_score[k] + O^4[j]$
    $k = argmax_{1 \le i \le K}(FBA\_score[i])$
    display action $a(t)$ corresponding to $k$ to the user; record clicks
    if user clicked $a(t)$ then
        update ($ANFIS_k$, reward ← 1)
    else
        continue
```

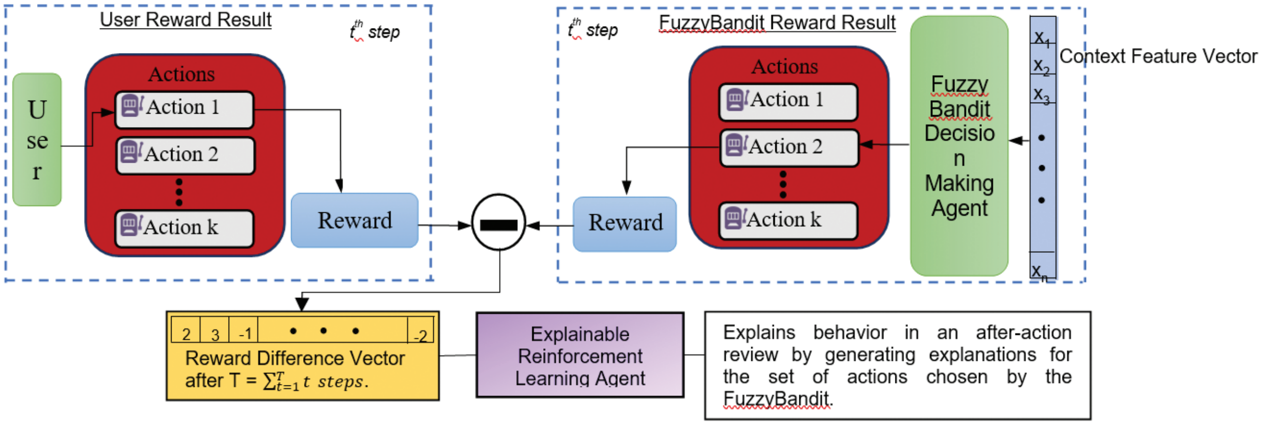
User Reward Result
t^{th} step
User
Actions
Action 1
Action 2
Action k
Reward
t^{th} step
FuzzyBandit Reward Result
Actions
Action 1
Action 2
Action k
Reward
Fuzzy Bandit Decision Making Agent
x_1
x_2
x_3
x_n
Context Feature Vector
2 3 -1 • • • -2
Reward Difference Vector after T = $\sum_{t=1}^{T} t$ steps.
Explainable Reinforcement Learning Agent
Explains behavior in an after-action review by generating explanations for the set of actions chosen by the FuzzyBandit.

Figure 2. Explainable AI model for FuzzyBandit.

whether the selected arm by the proposed model is the optimal arm or not and generates relevant feedback for network training. The feedback is the difference in the score calculated by the proposed FuzzyBandit model and the actual score using the Recursive Least Square Estimator (RLSE) method. In case the selected arm is the optimal arm, positive feedback is generated and back propagated to the selected arm. This is done to ensure that the next time the model is presented with the same user context feature vector, the previously selected arm *FBA_i*(*t*) score is higher than its peer arms, thereby exploiting the same action/choice as by the user in the past. Also, if the arm outputted by the FB-ATU is not the desired arm, negative feedback is not generated to minimize the score of the selected arm. This enables the model to explore a new action in the subsequent trial. Arm Tuning based parameters are trained using the derivation as proposed by Jang[22].

3.1 FuzzyBandit Ranking Algorithm

The FuzzyBandit Ranking Algorithm consists of three subsections. The first one is the *FBA_fuzzification* subsection computes fuzzy membership value associated with the user contextual feature vector for each FBA and stores it into the matrix, O^1 of dimension $n \times m$ where n is the number of features of the user contextual feature vector, and m is the number of membership functions associated with each feature. The second subsection, *FBA_integration,* first combines the membership of all features computed from the previous subsection by multiplying the membership values calculated in the matrix, O^1 and then normalizing the membership of each feature as each feature is represented with more than one membership function.Lastly, the *FBA_defuzzification* subsection returns the crisp output score *FBA_score* for each FBA. To calculate FBA_score, the normalized weights calculated in FBA_normalization subsection are multiplied with a function, F, and stored in a matrix O^4. F is a linear function of arm-rule based parameters and the user context feature vector. The same can be visualized in the defuzzification layer (as shown in Fig. 1) which multiplies the inference of normalized firing strength of each node and the first-order polynomial of user contextual feature vector and arm-rule based parameters. The score for each FBA shown in Fig. 2.

The summation of matrix, O^4 and stored in FBA_score. Each FBA is ranked in descending order of the FBA_score. The arm with the lowest rank is selected and the action corresponding to the selected arm is displayed to the user. If the user clicks on the displayed action, then the action selected by the model is correct and positive feedback is sent back i.e. exploiting the choice for the next iteration. If the selected arm is not clicked by the user, then no feedback is sent back which allows the proposed model to explore new actions in the next iteration.

3.2 Explainable AI

The proposed FuzzyBandit model incorporates explainable AI(XAI) by computing FuzzyBandit-User reward difference so that the model learns by itself without any human intervention and a user can also easily understand the decisions made by the model. This enables the proposed model to be transparent and perform autonomously. Thus, the model's decisions in the real-world environment can be trusted by the user, thereby allowing the model to be deployed at scale.

Figure 2 shows how the explanations are generated by the model using the FuzzyBandit-User (FB-U) reward difference. The FuzzyBandit model inputs ann-dimensional context vector $b_i(t) \epsilon R^d$ associated with each arm i at time t and yields a random reward $r_{a(t)}(t)$ with unknown distribution $\theta_i(t)$ for the corresponding action $a_i(t)$ chosen. If the user disapproves of the action chosen by the model or seeks an explanation as to why $a_i(t)$ is chosen, then the user chooses an action $\tilde{a}_i(t)$ preferred over the action $a_i(t)$ by the user and observes the reward, $\tilde{r}_{\tilde{a}_i(t)}(t)$. Let the optimal arm at time t is $a^*(t)=argmax_{1 \leq i \leq N} \{\theta_t(b_i(t))\}$. The FuzzyBandit-User (FB-U) reward difference vector can be computed as:

$$\delta\left(r, a_i(t), \tilde{a}_i(t)\right) = r_{a_i(t)}(t) - r_{\tilde{a}_i(t)}(t) | \{b_i(t)\}^T{}_{i=1}$$

$$= \theta_t\left(b_{a_i(t)}(t)\right) - \theta_t\left(b_{\tilde{a}_i(t)}(t)\right)$$

$$\text{FB-U} = \delta(r, a(t), \tilde{a}(t)) \; \forall \, t \in 1,2,..T$$

For t =1,2,..T steps, total reward difference R(T) can be calculated as:

$$R(T) = \sum_{t=1}^{T} r_{a_i(t)}(t) - r_{\tilde{a}_i(t)}(t) | \{b_i(t)\}^T{}_{i=1}$$

If the total reward difference, R(T) is greater than zero that means the total reward generated by the FuzzyBandit model is more than that of the reward generated by the user. A high positive magnitude of R(T) shows the decisions made by the proposed FuzzyBandit model are much higher reward yielding as compared to the decisions made by the user and thus enables humans to understand the decisions or predictions made by the proposed model. In order to quantify this and get deeper insight, we look into each component Δ_c(FB-U) which numerically can take either 0, -1, or +1. Logically, each component of FB-U signifies either a positive or negative reason for choosing action $a_i(t)$ over user-generated action, $\tilde{a}_i(t)$. Since the FB-U vector is computed over

T steps, it can become overwhelming for the human to visualize each action preference individually, so in order to simplify this we have computed two sets ofpositive and negative reasons i.e. FB-U$^+$ and FB-U$^-$. Mathematically,

$$\text{FB-U}^- = \sum_c I\left[\Delta_c\left(\delta\left(r, a(t), \tilde{a}(t)\right)\right) < 0\right]. \left|\Delta_c\left(\delta\left(r, a(t), \tilde{a}(t)\right)\right)\right|\right]$$

and

$$\text{FB-U}^+ = \sum_c I\left[\Delta_c\left(\delta\left(r, a(t), \tilde{a}(t)\right)\right) > 0\right]. \Delta_c\left(\delta\left(r, a(t), \tilde{a}(t)\right)\right)\right]$$

where I is the identity function.

We have computed the confidence score, α using FuzzyBandit-User reward difference, which signifies how much percentage the user can trust the FuzzyBandit model decisions in real-time settings. Alternatively, α represents the total percentage of the FuzzyBandit model's decisions (or actions chosen) better than the user's decisions. Similarly, β is the total percentage of the user's decisions better than that of the FuzzyBandit model. They can be computed as follows:

$$\alpha = \frac{\text{FB-U}^+}{\text{FB-U}^+ + |\text{FB-U}^-|}$$

$$\beta = \frac{\text{FB-U}^-}{\text{FB-U}^+ + |\text{FB-U}^-|}$$

In the above Eqn., α and β act as the trust factor for the user as they quantify that by how much value the FuzzyBandit model's decisions are better or worse than the user's decisions. This enables our proposed FuzzyBandit model to be deeply coupled with explainable AI to yield insight into complex model decisions and deployed autonomously without any human interventions.

4. EXPERIMENTAL RESULTS

In this paper, four publicly available datasets, namely the Forest Cover type dataset, Mushroom dataset, Statlog (Shuttle) dataset, and Adult Income from the UCI Machine Learning Repository[31], are being used. These datasets have been widely used in the literature as benchmark datasets to measure the performance of various contextual multi-arm bandit algorithms. The objective of this experiment is to test the suitability of the proposed FuzzyBandit model by combining various criteria for the datasets. In this paper, we considered nine criteria[39], namely

Table 1. Performance of contextual bandit models on the mushroom database for the optimal arm

Parameters	Uniform sampling	RMS	Dropout	BootRMS	ParamNoise	BBAlphaDiv	FuzzyBandit	Banditron
Recall	0.502	0.829	0.735	0.873	0.739	0.818	**0.851**	0.669
Specificity	0.504	0.756	0.695	0.826	0.708	0.834	**0.853**	0.668
Precision	0.521	0.785	0.722	0.843	0.731	0.841	**0.862**	0.684
Prevalence	0.518	0.518	0.518	0.518	0.518	0.518	**0.518**	0.518
F score	0.511	0.809	0.727	**0.856**	0.735	0.829	**0.855**	0.674
MCC	0.006	0.588	0.431	0.701	0.4469	0.652	**0.703**	0.337
FM	0.260	0.650	0.530	**0.733**	0.540	0.690	**0.730**	0.450
CSI	0.340	0.680	0.570	**0.751**	0.581	0.710	0.748	0.510
Overall accuracy	50.31	79.39	71.59	85.03	72.39	82.57	**85.18**	66.84

Table 2. Performance of contextual bandit models on the adult database for the optimal arm

Parameters	Uniform sampling	RMS	Dropout	BootRMS	ParamNoise	BBAlphaDiv	FuzzyBandit	Banditron
Recall	0.499	0.758	**0.773**	0.770	0.747	0.673	0.742	0.613
Specificity	0.504	0.756	**0.770**	0.768	0.745	0.679	0.729	0.611
Precision	0.754	0.905	**0.910**	**0.910**	**0.900**	0.870	0.893	0.828
Prevalence	0.753	0.753	0.753	0.753	0.753	0.753	0.753	0.753
F1 score	0.599	0.823	**0.836**	**0.834**	0.816	0.759	0.809	0.704
MCC	0.002	0.459	**0.487**	0.482	0.438	0.308	0.420	0.194
FM	0.374	0.682	**0.704**	0.701	0.672	0.586	0.663	0.508
CSI	0.429	0.702	**0.719**	0.715	0.690	0.609	0.681	0.544
Overall accuracy	50.02	75.74	**77.2**	**76.93**	74.67	67.48	**73.87**	61.23

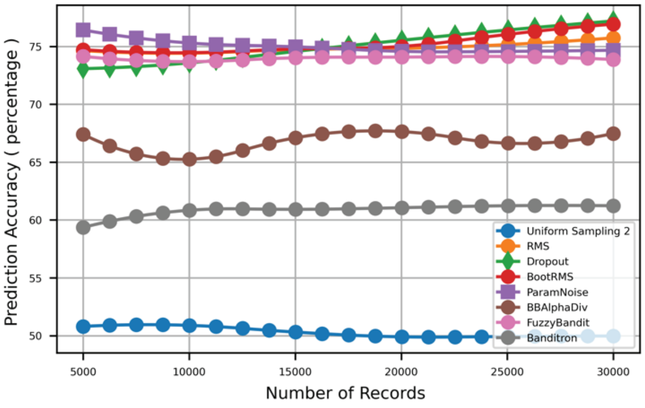

Figure 3. Prediction accuracy for adult dataset.

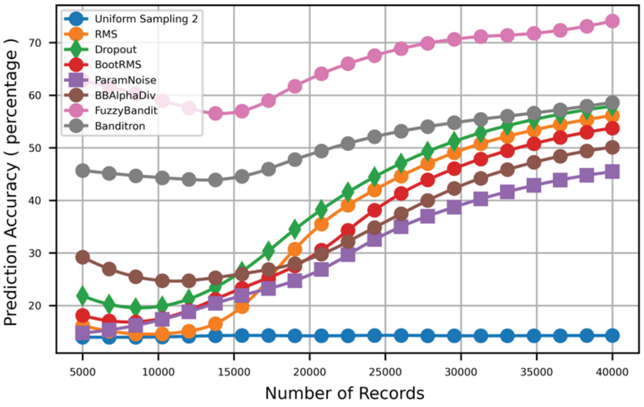

Figure 4. Prediction accuracy for covertype dataset.

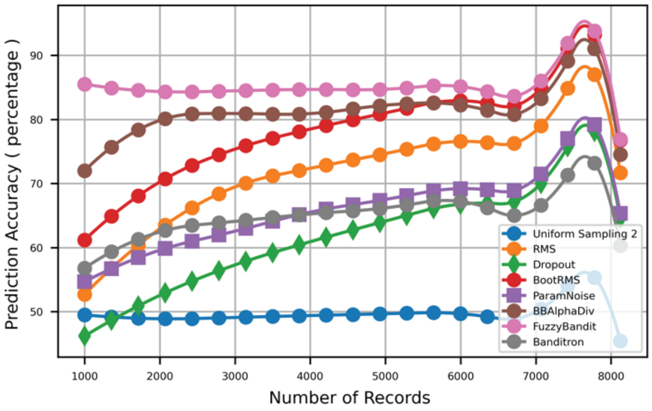

Figure 5. Prediction accuracy for mushroom dataset.

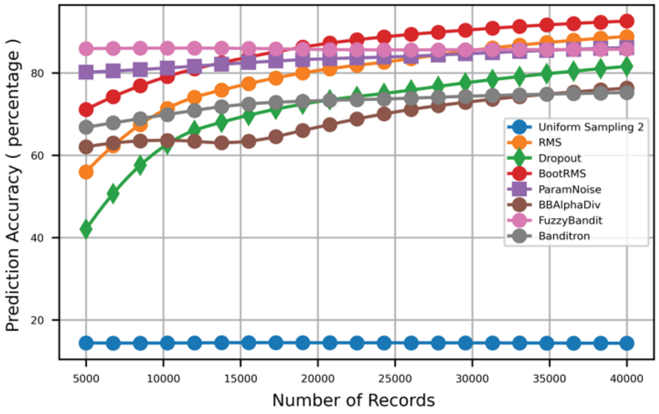

Figure 6. Prediction accuracy for statlog dataset.

recall, specificity, precision, prevalence, F1 score, Matthews correlation coefficient (MCC), Fowlkes–Mallows index (FM), critical success index (CSI), and accuracy as summarized in Table 1 and Table 2.These criteria[39] are important as the number of observations for each arm/ choice varies significantly in the dataset, and thereby, relying alone on accuracy can be misleading. For e.g., in the mushroom dataset, there are 3916 instances where the mushroom is edible and 4084 instances where the mushroom is poisonous, and the algorithm has to predict out of the two choices, i.e., mushroom is either edible or poisonous. A particular algorithm/model might predict all the observations as poisonous, giving an overall accuracy of 51 %, but in more detail, the algorithm has 100 % sensitivity for the poisonous class but a 0 % sensitivity for the edible class.

All the experiments are conducted in online settings, where the context is fed into the model and action is recorded. The model is then updated with the feedback observed. We have tested the performance of the proposed FuzzyBandit model on various criteria with the existing seven contextual bandit models. The first one is the uniform sampling model, a random policy model that ignores the context vector and chooses an arm randomly for each iteration. The second model is the Banditron model, which uses a simple perception model for exploitation and epsilon greedy as an exploration policy for each iteration. The third model is the Dropout model, a variation of the Neural Model (which uses each arm as a neural model for exploiting the best arm and epsilon greedy policy for exploration) and uses the Dropout neural network to predict the best arm in each trial. Next is the RMS model, which trains a neural network and chooses the action based on the highest score predicted, i.e., acts greedily for the current context. BootRMS model uses Bootstrap[36] to offer significant performance gain with respect to its parent RMS model.

The remaining models, namely BBAlphadiv[35] and Param Noise[34], are non-linear models and use Thompson Sampling for exploitation.

4.1 Adult Income Dataset

In Figure 3, the curves for the five models, namely the RMS, the BootRMS, the Dropout, the ParamNoise, and the FuzzyBandit model, overlap where The RMS model and the dropout model take the lead in the latter part of the experiment with an accuracy of 73-77 %. The FuzzyBandit mode with 5 % exploration achieved an accuracy of about 73-75 % and performed better than the Banditron and the BBAAlphadiv model.

4.2 Cover Type Dataset

An interesting trend in the curves of all the models can be seen in Fig. 4. The accuracy for all the models initially dips and then increases as the experiment progresses. The FuzzyBandit model emerges as a clear winner and achieves an accuracy of about 65-70 %, beating all other models by a significant margin. Surprisingly, the banditron model is the second-best-performing model, with an accuracy of around 48-52 %.

4.3 Mushroom Dataset

The experimental results on the Mushroom dataset in Fig. 5 show that all the curves of different models follow a similar trend. They reach a peak and then start to fall as the experiment progresses. However, the FuzzyBandit model adapts better to new input than its other neural model counterpart and flattens. The FuzzyBandit model achieves a maximum accuracy of 93 % with an average accuracy of about 85-87 %, outperforming the other models except for the BootRMS model, which performs almost similar to the FuzzyBandit model.

4.4 Statlog Dataset

In Fig. 6, it can be seen that the curves of all the models show an increasing trend as the experiment progresses.Initially, the FuzzyBandit model takes the lead by outperforming all other models, but as the experiment progresses, the accuracy of the FuzzyBandit model remains the same at around 83 %, while the BootRMS model takes the lead. The ParamNoise model performs at par with the FuzzyBandit model with an accuracy of 80-83 %.

5. CONCLUSION

A new FuzzyBandit model for the contextual multi-arm bandit problem has been proposed. Each arm in the FuzzyBandit model mimics ANFIS independently by adjusting arm tuning and arm-rules based parameters according to user choice. The model can easily tune all its parameters by back propagation and can be successfully trained in both offline and online settings for real-time applications where the rewards and data distribution can be dynamic. A comparative study of the contextual bandit's models on nine criteria has been proposed. It is observed that none of the models has outperformed the other remaining models on different criteria in all four datasets. However, the FuzzyBandit model has shown promising results for all the datasets.

REFERENCES

1. Agrawal, S. & Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. *In* Conference on learning theory, 39.1-39.26, 2012. doi:10.48550/arXiv.1111.1797
2. Allesiardo, R.; Féraud, R. & Bouneffouf, D. A neural networks committee for the contextual bandit problem. *In* Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, pp. 374-381, 2014. doi: 10.1007/978-3-319-12637-1_47
3. Auer, P.; Cesa-Bianchi, N.; Freund, Y. & Schapire, R.E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 2002, **32**(1), 48-77. doi:10.1137/S0097539701398375
4. Auer, P.; Cesa-Bianchi, N.; Freund, Y. & Schapire, R.E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *In* Proceedings of IEEE 36th Annual Foundations of Computer Science, 322-331, 1995. doi: 10.1109/SFCS.1995.492488
5. Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1993, **25**(3/4), 285-294. doi:10.1093/biomet/25.3-4.285

6. Liu, B.; Yu, T.; Lane, I. & Mengshoel, O. Customized non-linear bandits for online response selection in neural conversation models. *In* Proceedings of the AAAI Conference on Artificial Intelligence, 2018. doi:10.48550/arXiv.1711.08493
7. Chapelle, O. & Li, L. An empirical evaluation of thompson sampling. *Adv. Neural Inf. Process. Syst.,* 2011, **24**, 2249-2257. doi:10.48550/arXiv.2101.08699
8. Chu, W.; Li, L., Reyzin, L. & Schapire, R. Contextual bandits with linear payoff functions. *In* Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 208-214, 2011. doi:10.48550/arXiv.1209.3352
9. Crammer, K. & Gentile, C. Multiclass classification with bandit feedback using adaptive regularization. *Mach. Learn.*, 2013, **90**(3), 347-383. doi:10.1007/s10994-012-5319-1
10. Vermorel, J. & Mohri, M. Multi-armed bandit algorithms and empirical evaluation. *In* European conference on machine learning, 437-448, 2005. doi:10.1007/11564096_42.
11. Dudik, M.; Hsu, D.; Kale, S.; Karampatziakis, N.; Langford, J.; Reyzin, L. & Zhang, T. Efficient optimal learning for contextual bandits, 2011, arXiv preprint. doi:10.48550/arXiv.1106.2369
12. Lai, T.L.; Lavori, P.W. & Tsang, K.W. Adaptive design of confirmatory trials: Advances and challenges. *Contemp. Clin. Trials*, 2015, **45**, 93-102. doi: 10.1016/j.cct.2015.09.005
13. Zhou, Q.; Zhang, X.; Xu, J. & Liang, B. Large-scale bandit approaches for recommender systems. *In* International Conference on Neural Information Processing, 811-821, 2017. doi:10.1007/978-3-319-70087-8_83
14. Bouneffouf, D.; Rish, I. & Cecchi, G.A. Bandit models of human behavior: Reward processing in mental disorders. *In* International Conference on Artificial General Intelligence, 237-248, 2017. doi: 10.1007/978-3-319-63703-7_22
15. Kakade, S.M.; Shalev-Shwartz, S. & Tewari, A. Efficient bandit algorithms for online multiclass prediction. *In* Proceedings of the 25th international conference on Machine learning, 440-447, 2008. doi:10.1145/1390156.1390218
16. Chen, D.; Cao, X.; Wang, L.; Wen, F. & Sun, J. Bayesian face revisited: A joint formulation. *In* European conference on computer vision, 566-579, 2012. doi: 10.1007/978-3-642-33712-3_41
17. Lai, T.L. & Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 1985, **6**(1), 4-22. doi:10.1016/0196-8858(85)90002-8
18. Langford, J. & Zhang, T. Epoch-Greedy algorithm for multi-armed bandits with side information. *Adv. Neural Inf. Process. Syst.*, 2007, **20**(1).
19. Sugeno, M. & Kang, G.T. Structure identification of fuzzy model. *Fuzzy Sets Syst.*, 1988, **28**(1), 15-33. doi: 10.1016/0165-0114(88)90113-3
20. Li, L.; Chu, W.; Langford, J. & Schapire, R.E. A contextual-bandit approach to personalized news article recommendation. *In* Proceedings of the 19th international conference on World wide web, 661-670, 2010. doi: 10.1145/1772690.1772758
21. Tokic, M. Adaptive ε-greedy exploration in reinforcement learning based on value differences. *In* Annual Conference on Artificial Intelligence, 203-210, 2010. doi: 10.1007/978-3-642-16111-7_23
22. Takagi, T. & Sugeno, M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern.*, 1985, **1**, 116-132. doi:10.1109/TSMC.1985.6313399
23. Jang, J.S. ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.*, 1993, **23**(3), 665-685. doi:10.1109/21.256541
24. Li, L.; Chu, W., Langford, J. & Schapire, R.E. A contextual-bandit approach to personalized news article recommendation. *In* Proceedings of the 19th international conference on World wide web, 2010, 661-670. doi:10.1145/1772690.1772758
25. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 2014, **15**(1), 1929-1958. doi:10.5555/2627435.2670313
26. Robbins, H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.*, 1952, **58**(5), 527-535. doi:10.1090/S0002-9904-1952-09620-8
27. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 1958, **65**(6), 386. doi:10.1037/h0042519
28. Riquelme, C.; Tucker, G. & Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. 2018, arXiv preprint. doi:10.48550/arXiv.1802.09127
29. Auer, P.; Cesa-Bianchi, N. & Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.,* 2002, **47**(2), 235-256. doi:10.1023/A:1013689704352
30. Auer, P. & Cesa-Bianchi, N. Online learning with malicious noise and the closure algorithm. *Ann. Math. Artif. Intell.,* 1998, **23**(1), 83-99. doi: 10.1023/A:1018926100936
31. Dua, D. & Graff, C. UCI Mach. Learn. Irvine, CA: Univ. California, *Sch. Inf. Comput. Sci.,* 2019 http://archive.ics.uci.edu/ml.
32. Gai, Y.; Krishnamachari, B. & Jain, R. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. *In* 2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN), 1-9, 2010. doi:10.1109/DYSPAN.2010.5457871
33. Lai, L.; Jiang, H. & Poor, H.V. Medium access in cognitive radio networks: A competitive multi-armed

bandit framework. *In* 2008 42nd Asilomar Conference on Signals, Systems and Computers, 98-102, 2008. doi:10.1109/ACSSC.2008.5074491

34. Plappert, M.; Houthooft, R.; Dhariwal, P.; Sidor, S.; Chen, R.Y.; Chen, X.; Asfour, T.; Abbeel, P. & Andrychowicz, M. Parameter space noise for exploration, 2017, arXiv preprint. doi:10.48550/arXiv.1706.01905.
35. Hernandez-Lobato, J.; Li, Y.; Rowland, M.; Bui, T.; Hernández-Lobato, D. & Turner, R. Black-box alpha divergence minimization. *In* International Conference on Machine Learning, 1511-1520, 2016. doi: 10.17863/CAM.9689
36. Efron, B. The jackknife, the bootstrap and other resampling plans. *Soc. Ind. Appl. Math.,* 1982. doi:10.1137/1.9780898716961
37. Tang, L.; Jiang, Y.; Li, L. & Li, T.Ensemble contextual bandits for personalized recommendation. *In* Proceedings of the 8th ACM Conference on Recommender Systems, 73-80, 2014. doi: 10.1145/2645710.2645751
38. Luo, X.; Zhou, M.; Li, S.; You, Z.; Xia, Y. & Zhu, Q. A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method. *IEEE Trans. Neural Networks Learn. Syst.*, 2015, **27**(3), 579-592. doi: 10.1109/TNNLS.2015.2417194
39. Kumari, T.; Sharma, R. & Bedi, P. A contextual-bandit approach for multifaceted reciprocal recommendations in online dating. *J. Intell. Inf. Syst.,* 2022, 1-27. doi:10.1007/s10844-022-00708-6
40. Iovine, A.; Lops, P.; Narducci, F.; de Gemmis, M. & Semeraro, G. An empirical evaluation of active learning strategies for profile elicitation in a conversational recommender system. *J. Intell. Inf. Syst.,* 2022, 1-26. doi:10.1007/s10844-021-00683-4
41. Panda, D.K. & Ray, S. Approaches and algorithms to mitigate cold start problems in recommender systems: A systematic literature review, *J. Intell. Inf. Syst.,* 2022, **59**(2), 341-366. doi:10.1007/s10844-022-00698-5.

CONTRIBUTORS

Mr Nipun Bansal is an Assistant Professor at the Department of Computer Science and Engineering at Delhi Technological University. He obtained his Master’s in Technology from IIIT Delhi.
In the current study, he designed and implemented the proposed FuzzyBandit model in the manuscript. He is responsible for drafting, editing, and reviewing the complete manuscript.

Dr Manju Bala is working as an Assistant Professor in the Department of Computer Science at I.P. College for Women, Delhi University. She obtained her Doctoral degree in computer science and engineering at Jawaharlal Nehru University.
In the current study, she has provided guidance and mentorship to successfully implement the FuzzyBandit model using XAI.

Dr Kapil Sharma is working as a Professor in the Department of Information Technology at Delhi Technological University. He obtained his Doctorate degree in Computer Science and Engineering under the Faculty of Engineering and Technology at the M.D. University, Rohtak (Haryana), India.
In the current study, he has provided guidance and mentorship to design and implement a novel personalized contextual multi-arm bandits based model using Explainable AI.