Predicting early user churn in a public digital weight loss intervention

Robert Jakob
Centre for Digital Health Interventions, ETH Zurich, Zurich, Switzerland
rjakob@ethz.ch

Nils Lepper
Centre for Digital Health Interventions, ETH Zurich, Zurich, Switzerland
nlepper@student.ethz.ch

Elgar Fleisch
Centre for Digital Health Interventions, ETH Zurich, Zurich, Switzerland; University of St. Gallen, St. Gallen, Switzerland
efleisch@ethz.ch

Tobias Kowatsch
Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland; School of Medicine, University of St. Gallen, St. Gallen, Switzerland
tobias.kowatsch@uzh.ch

ABSTRACT

Digital health interventions (DHIs) offer promising solutions to the rising global challenges of noncommunicable diseases by promoting behavior change, improving health outcomes, and reducing healthcare costs. However, high churn rates are a concern with DHIs, with many users disengaging before achieving desired outcomes. Churn prediction can help DHI providers identify and retain at-risk users, enhancing the efficacy of DHIs. We analyzed churn prediction models for a weight loss app using various machine learning algorithms on data from 1,283 users and 310,845 event logs. The best-performing model, a random forest model that only used daily login counts, achieved an F1 score of 0.87 on day 7 and identified an average of 93% of churned users during the week-long trial. Notably, higher-dimensional models performed better at low false positive rate thresholds. Our findings suggest that user churn can be forecasted using engagement data, aiding in timely personalized strategies and better health results.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Learning paradigms**; • **Supervised learning**; • **Supervised learning by classification;**; • **Human-centered computing**; • **Ubiquitous and mobile computing**; • **Empirical studies in ubiquitous and mobile computing;**; • **Applied computing**; • **Life and medical sciences**; • **Health informatics**;

KEYWORDS

machine learning, churn, dropout, attrition, mHealth, digital health

ACM Reference Format:
Robert Jakob, Nils Lepper, Elgar Fleisch, and Tobias Kowatsch. 2024. Predicting early user churn in a public digital weight loss intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642321

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642321

1 INTRODUCTION

Digital health interventions (DHIs) demonstrate the potential to assist patients and healthcare systems in tackling the global increase and financial impact of noncommunicable diseases (e.g., cardiovascular diseases, diabetes, or mental health conditions), the leading cause of death and disability worldwide [23, 33, 63, 83, 86]. In particular, mobile health (mHealth) apps have emerged as versatile tools for promoting behavior changes among patients, improving health outcomes, and reducing healthcare costs due to the widespread availability of smartphones [16, 55, 71, 82]. Weight loss apps, a subset of mHealth apps, play a vital role in preventing noncommunicable diseases by promoting healthy lifestyles in the general population [24, 38, 80]. These apps function as personal health coaches, offering features like dietary tracking, exercise regimens, social support, and educational content [24, 38, 80]. Recent studies have shown that regular use of weight loss apps can lead to effective weight reduction, prevent non-communicable diseases, and offer a cost-effective and accessible alternative to traditional programs [38, 80].

However, sustaining user engagement over an extended period poses a significant challenge for weight loss apps and DHIs. The occurrence of user churn - also referred to as dropout, non-adherence, disengagement, or attrition - where users stop using apps prematurely hampers the potential long-term health benefits these tools could provide [17, 28, 29, 61]. Prior research underscores that churn is a substantial concern in DHIs, with observational studies recording a higher dropout rate of 49% compared to 40% in more controlled study settings [47]. The issue of high churn rates also extends to weight loss apps [24, 29]. A recent review revealed that average adherence rates in weight loss apps (n = 9, 49.1%), determined as the ratio between intended and actual use, were comparably lower than the average adherence rate of other app domains, e.g., diabetes, cancer, or cardiovascular disease management apps (n = 88, 56.7%). This is especially the case for publicly available weight loss apps (n = 4, 42.2%) as compared to those used exclusively in controlled study settings (n = 5, 54.7%) [29]. A study by Baumel et al. (2019) examining retention rates across 59 DHIs also revealed a significant drop in user retention within the initial seven days of app usage [4]. After the first seven days, fewer than an average of 10% of users continued to log in to the DHIs daily. This highlights notably high churn rates during the initial week of DHI usage.

A small but growing body of research studies indicates that churn prediction can aid mHealth app providers in identifying users at risk of disengaging and in delivering personalized interventions to retain them, thereby amplifying their effectiveness [7, 20, 39, 54, 62, 78]. However, these studies vary widely regarding methodologies, such as the choice of mHealth apps, machine learning (ML) algorithms, feature selection, and model dimensionality. Notably, only one of these studies considered a publicly available weight loss intervention [39]. Most research predicts churn after the first week of usage when most users already churned [4]. Thus, it remains unclear which features, ML algorithms, and model dimensionalities are most promising for early churn prediction in publicly available DHIs.

Another significant gap in the literature is the absence of studies examining the number of users who reengage with DHIs once correctly predicted by churn prediction models. This information is crucial for assessing the potential of in-app churn interventions and timely adaptations of persuasive and behavior-change systems to retain users.

Our study seeks to fill these research gaps. We evaluate churn prediction models for a publicly available and subscription-based weight loss app in the first seven days of user interaction. Our evaluation employs a variety of ML algorithms and dimensionality configurations, utilizing data from 1,283 users and 310,845 event logs. By analyzing user reengagement following accurate churn predictions, we further aim to assess intervention potential.

2 RELATED WORK

2.1 Churn prediction

Churn prediction is a specialized domain within data analytics that uses ML algorithms to detect users likely to discontinue a product or service [21]. It has found application across numerous industries, notably telecommunications [50], financial services [73], gaming [36], and e-commerce [25]. Recently, there's been a surge in interest in applying churn prediction to mHealth apps. A small but growing number of studies indicate that ML algorithms can accurately predict churn in mHealth apps, underscoring the viability of such an approach [7, 20, 39, 54, 62, 69, 78]. In the context of mHealth apps, churn is generally defined as user inactivity over a certain period or the event of uninstalling or unsubscribing from the app [7, 39, 39, 54, 62, 69, 78]. The prediction process involves data preprocessing methods such as data cleaning, normalization, and transformation to ensure data quality and reliability [1, 15, 79]. Subsequently, feature selection takes place. In this context, features are specific user attributes or data points that can be used to predict churn. Common features in churn prediction are self-reported user data (e.g., personal goals, dietary preferences), measured user data (e.g., step counts from integrated fitness trackers), and app engagement data (e.g., app logins, session durations), encompassing both static and time-series data [1, 12, 51, 81]. Imbalanced data poses a common challenge, where the number of churned users in mobile apps is substantially higher than retained users. To address this, techniques such as undersampling, oversampling, or synthetic minority over-sampling technique (SMOTE) are commonly employed [13, 15]. Various ML algorithms like Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Gradient Boosting Machines (e.g., XGBoost), neural networks, sequence models, or ensemble methods are generally used for churn prediction [1, 12, 15, 25, 31, 50, 51, 73, 79, 81]. When working with longitudinal data, time series, and survival analysis are also used in churn prediction [6, 42, 54, 69].

2.2 Review of churn prediction results in previous mHealth app studies

Regarding individual study results, Trinh et al. (2018) identified churned users among a total of 61 users in the second week of using the virtual agent-based mHealth app "Tanya" with 90% accuracy (F1 = 0.84) using gradient boosting and first-week session frequency features [78]. Pedersen et al. (2019) demonstrated an 86% accuracy rate (AUC = 0.92) in identifying user churn among 2,684 patients, defined as four weeks of user inactivity, in an eHealth platform for chronic lifestyle diseases using a random forest model and 11 predictor variables including user demographic and app engagement data [62]. Notably, users who churned within the initial 14 days were not considered in this study. Yet, the researchers did hint at the potential of a future study focusing on these "very early dropouts" [62]. Ganju et al. (2021) achieved a 93% accuracy rate in predicting app uninstalls among 45,000 users in a family health app using a random forest model and user app engagement data (e.g., engagement times, session counts), unfortunately, without clearly reporting the observation period [20]. Kwon et al. (2021) employed a recurrent neural network structure to predict churn after 16 weeks, defined as users requesting a refund, in a holistic lifestyle health app with 1,868 users, achieving 89% accuracy (F1 = 0.89, AUC = 0.89) using user demographic data, app engagement data, and text data [39]. Schleicher et al. (2022) used time-series analysis and XGBoost to identify different phases of user engagement and predict periods of inactivity across 3,177 users. Utilizing only time-series data from a user questionnaire, this method achieved a mean accuracy of 77% in predicting engagement in the later "stable use phase" of the intervention [69]. Olaniyi et al. (2022) applied probabilistic and survival analysis methods to study churn among 58,195 users of a mobile health app for frontline healthcare workers using various app engagement data, with left-truncated and right-censored models performing the best (Integrated Brier Score = 0.09) [54]. Lastly, Bricker et al. (2023) predicted early dropout, defined as users who only log into the app in the first week, in the two smoking cessation mHealth apps "iCanQuit" (1,069 users) and "QuitGuide" (1064 users). Their logistic regression model, using only the first 7 days of login count data, predicted churn with an AUC of 0.94 and 0.88 for "iCanQuit" and "QuitGuide", respectively [7].

In conclusion, while these studies provide compelling evidence that churn in mHealth apps can be predicted using ML algorithms, there's a notable diversity in their methodologies. This includes variations in the selection of mHealth apps, ML algorithms, feature choices, and model complexity. Importantly, only one study focused on a publicly accessible weight loss intervention [39]. A majority of the research predicts churn after the initial week of app usage, a period when many users might have already discontinued use with some studies even excluding users who churn within the first few

weeks [4, 62]. Another significant gap in the current literature is the lack of research on the number of users who reengage with the app after a correct churn prediction. This data is vital for assessing the real-world efficacy of in-app churn interventions.

For the practical application of churn prediction models in informing targeted churn interventions, models with fewer features (or low dimensionality) are likely preferable. Such low-dimensional models are more interpretable, easier to deploy, and more adaptive to changes in the app and data infrastructure. In this context, the high churn prediction performance of Bricker et al. (2023) using only daily user login counts as features stands out as promising [7]. Yet, the potential enhancement in model performance upon the inclusion of more features, thereby expanding model dimensionality, remains an open question.

2.3 Relevance of churn prediction in digital health interventions for HCI research

Churn prediction involves the analysis of user behavior patterns and factors that predict churn. Insights from these analyses can support Human-Computer Interaction (HCI) research's ongoing focus on enhancing the effectiveness of persuasive and behavior-change systems, which have received substantial attention in the HCI literature in digital health interventions [2, 9, 45, 49, 64, 67, 75], and specifically in interventions targeting physical activity [35, 40, 61, 85], diet [19, 56], and weight loss [3, 14, 26, 66]. A significant trend in the HCI literature is the shift from generic "one-size-fits-all" systems to personalized, context-sensitive, and adaptive systems [32, 56–59, 84].

Prior HCI research suggests that personalizing persuasive and behavior change systems to individual user characteristics enhances user engagement and effectiveness of these systems [30, 32, 37, 43, 52, 57–60, 84, 85]. The concept of 'persuasion profiles' emphasizes the need for tailored interventions based on individual susceptibilities to different persuasive strategies [18, 32, 40]. However, information on the user's characteristics, preferences, and responses to match persuasion strategies is not always attainable in real-world settings. Given that disengagement relates to a persuasion profile mismatch, churn prediction can act as an early indicator to adapt persuasion strategies as an intervention mechanism. Such interventions could also include a preceding questionnaire to collect user information that allows for adequate persuasion profile mapping.

Regarding context-sensitive and adaptive systems, prior HCI research concludes that persuasive and behavior change systems need to adapt their strategies to the user's contextual factors [34, 61, 64] and various stages of behavior change in accordance with the transtheoretical model to effectively support individuals at each stage [37, 59, 65]. A notable point in DHIs is that disengagement can also indicate "Happy Abandonment", reflecting the successful accomplishment of a DHI's intended goal, often sustained behavior change [10, 11, 53, 70]. One challenge in adapting systems to different stages in the user's behavior change journey is detecting the point of stage transitioning. Likewise, it is difficult to distinguish happy abandonment from unintended abandonment since this information is hardly attainable once users have fully abandoned a system. Churn prediction can potentially help detect disengagement and trigger questionnaires that collect information on the user's reasons for disengagement, particularly on their attitude and behavior change, before access to these users is lost [70]. Besides collecting relevant information on which aspects of the system or the user's context are related to disengagement, this process could aid in distinguishing between happy and unintended disengagement.

In essence, churn prediction has the potential to provide actionable insights that can enhance HCI research, particularly in areas related to adaptive persuasive and behavior change systems. By detecting users that are about to churn, these models may inform persuasive and behavior change systems to adapt and intervene timelier and, thus, more effectively.

To exploit this utility, however, it is crucial that churn prediction models correctly classify users at risk before they fully disengage, for users to receive passive system adaptations or active interventions. This potential, however, has not been evaluated in previous studies on churn prediction in DHIs. It remains unclear if churn prediction can effectively detect a user who is about to disengage or only detect fully disengaged users more quickly.

3 OBJECTIVES

Given the increasing adoption of weight loss apps by the general public and the growing relevance of churn prediction for the prevention of churn in mHealth apps, we aim to develop and evaluate ML models that predict early churn on a real-world dataset of the public digital weight loss intervention WayBetter.

Unlike previous churn prediction studies in mHealth apps, ours is the first to explore the prediction of churn, defined as the act of unsubscribing from the intervention within or following a 7-day trial period. The subscription model offers two notable aspects for studying early user churn: First, prior to receiving app access, users need to agree to a paid 6-months subscription that activates automatically after the 7-day trial, also requiring users to enter their credit card information in advance. This procedure leads to a natural preselection of users with a sincere intention to use the app compared to studies in free publicly available apps. Second, the necessity for users to actively unsubscribe, offers a clear churn indicator, providing a more accurate measure than relying on user inactivity after some predefined period.

Given particularly high churn rates of DHIs in the first week of usage, we evaluate churn prediction models after each of the first seven days of user interaction to understand how prediction performance develops over time. For the prediction, we use commonly applied ML algorithms, including Logistic Regression (LR), Decision Tree (DT), Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), Artificial Neural Network (NN), and an Ensemble method (ENS) to allow for comparison with previous studies. Furthermore, we are the first to assess the performance of Bayesian Logistic Regression (BLR) for churn prediction in DHIs. BLR and approximate Bayesian algorithms, in general, can also obtain measures of confidence for the prediction and are thus well suited to inform decisions in just-in-time adaptive interventions (JITAIs), particularly those applying reinforcement learning (RL). JITAIs applying RL recently gained attention for their ability to explore and exploit sequential intervention decisions automatically [44, 72, 77]. Comparing churn prediction performances of BLR with

commonly applied methods will help inform the potential of these models for churn prediction and prevention in self-adapting RL models.

We further investigate which features excel in predicting churn and how varying model dimensionalities influence the performance of churn prediction models. As a baseline, we adopt a low-dimensional model informed by Bricker et al. (2023) that predicts churn based on the number of app logins per day of the first seven days of user interaction [7]. Furthermore, we utilize medium and high-dimensional models with additional user-related and app-engagement features to explore to which extent model performance can be improved with additional features. For our best-performing model, we also evaluate which features are most important for churn prediction performance.

Finally, to estimate the potential for churn interventions following churn prediction within the 7-day trial period, we are also the first to compare churn prediction results on days 1-6 with subsequent user in-app activity. Thus, we assess how many users reengage with the intervention after being correctly categorized as churned users before the trial period runs out. These findings will provide a clearer understanding of how churn prediction can guide timely adaptations in persuasive and behavior-change systems. In summary, our research questions are as follows:

- **RQ1:** Which ML algorithms are most effective in predicting early churn on each of the first seven days of app usage?
- **RQ2:** How do low-dimensional churn prediction models using only daily logins compare against higher-dimensional models with more available features?
- **RQ3:** Which features have the highest predictive performance for predicting churn in a weight loss app?
- **RQ4:** How many users reengage with the app after a correct churn prediction?

4 METHODS

4.1 Dataset and definition of churn

This study analyzed an anonymized dataset provided by WayBetter Inc., from their publicly available mHealth app "WayBetter," which is accessible on both iOS and Android platforms. Apart from providing the anonymized dataset, WayBetter Inc. was not involved in the conduct of this study. The authors declare no conflicts of interest pertaining to the company. Due to the anonymous nature of the dataset, this study has been exempted from ethics approval by the research team's University Ethics Committee.

The WayBetter app is intended to facilitate weight loss by employing an integrative approach that merges motivational strategies, expert-driven coaching, and community support. The premise of the application is rooted in behavioral economics, particularly the concept of commitment contracts, which encourage users to place a monetary wager on achieving specific behavioral health goals. The app's intervention components have previously demonstrated efficacy in encouraging weight loss and have been associated with a clinically relevant increase in step counts in related interventions [8, 41].

The application offers a dual-component system to maintain adherence and engagement. Firstly, a gaming component provides users with access to curated games in three categories (mindset, nutrition, and fitness) that have a duration between one to six weeks. For example, popular games include journaling health-related activities (mindset), meal tracking (nutrition), achieving step goals, and working out (fitness). All players who have placed wagers and achieved the game's goal are declared "winners" and split the total sum of money equally, so they receive a full refund of their wager plus extra profit.

To move up game levels, users need to successfully complete at least one game per category. This is designed to offer users a customizable experience tailored to their individual lifestyles and preferences. Secondly, the application leverages community features that enable social interactions with other users for group challenges, the exchange of progress updates, and the provision of mutual support to foster collective motivation. WayBetter also incorporates interoperability with various fitness trackers and health apps to provide a holistic view of an individual's health metrics. Alongside users discovering the app through app store discovery and word-of-mouth (organic users), WayBetter also runs targeted social media advertising campaigns, e.g., via Facebook, to acquire additional users (paid users).

To access the WayBetter app, users are currently required to opt for a six-month subscription priced at USD 69. During the initial 7-day trial phase, users have the option to unsubscribe without incurring any charges. In this study, users were categorized as 'churned users' if:

- They cancel their subscription within the 7-day trial period.
- They request a refund within the first 14 days of paid subscription.
- Their subscription fee payment fails.

Conversely, users who maintain their subscriptions beyond these conditions are categorized as 'retained users'.

The dataset comprises historical data from 1,293 users who initiated a 7-day trial during the week of April 3rd to April 9th, 2023. With the aim of predicting churn during the trial phase, we established a 14-day observation window from April 3rd to April 16th, 2023, also ensuring no app updates occurred during this time. The dataset contains static attributes (e.g., user weight) and clickstream data (e.g., app login or access of a certain feature), totaling 416,262 app event log entries across 180 event types (e.g., user views his activity plan).

4.2 Data cleaning, manipulation, and analyses

We excluded six app event types that were directly related to the event of unsubscribing from the app, e.g., user navigating to the membership cancellation or delete account screen, or user providing the primary reason for canceling membership. Users without a single in-app session timestamp were excluded from the dataset (n = 10). We further excluded duplicate event logs (n = 105,257). Handling outliers, in seven cases we imputed user weight as values were either 0 (n = 5) or above 1000 lbs (n = 2). Based on event logs, user sessions and session lengths were derived. We also constructed event log counts in each session. Sessions with a time duration of zero seconds were excluded (n = 222). Also, sessions that ended before the subscription start date (n = 107) or started after the trial end date (n = 320) were excluded. The resulting dataset used for

Table 1: Number of available features used in LDMs (Low-Dimensional-Models), MDMs (Medium-Dimensional-Models), and HDMs (High-Dimensional-Models) on each trial day for churn prediction.

	Number of available features							Feature availability		
Feature name	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**	**Day 7**	**LDM**	**MDM**	**HDM**
(1) Number of app logins per day	1	2	3	4	5	6	7	x	x	x
(2) Minutes spent in the app per day	1	2	3	4	5	6	7		x	x
(3) Number of unique app events per day	1	2	3	4	5	6	7		x	x
(4) Number of total app events per day	1	2	3	4	5	6	7		x	x
(5) Weekday of install	6	6	6	6	6	6	6		x	x
(6) User weight	1	1	1	1	1	1	1		x	x
(7) User acquisition type (paid / organic)	1	1	1	1	1	1	1		x	x
(8) Minutes spent in each app event per day	174	348	522	696	870	1044	1218			x

LDM = Low-Dimensional-Model; MDMs = Medium-Dimensional-Model, HDM = High-Dimensional-Model

analyses consisted of 1,283 users with a total of 8,255 sessions and 310,845 event logs across 174 unique app events.

Based on this refined dataset, we constructed the following features for modeling: *(1) Number of app logins per day* (numerical), *(2) minutes spent in the app per day* (numerical*), (3) number of unique app events per day* (numerical), *(4) number of total app events per day* (numerical), *(8) minutes spent in each app event type per day* (numerical). We also constructed the user's *(5) weekday of install* (categorical: Monday - Sunday) as previous studies suggest that downloads on weekends are related to churn [27, 29]. Furthermore, we included *(6) user weight* (numerical), as previous research indicates that greater disease severity, and thus likely greater weight, is associated with increased churn rates [5, 29, 74]. Finally, we included *(7) user acquisition type* (categorical: Paid or Organic). Users acquired through paid channels, such as social media ads, reportedly exhibit different retention patterns than organic users who find the app via app store searches or word-of-mouth [87].

4.3 Model training and evaluation

We preprocessed available features by one-hot encoding categorical features and normalizing numerical data by performing square root scaling for right-skewed data and standard scaling for non-right-skewed data. We randomly split 80% of the dataset into a training set (n = 1026) and the remaining 20% into a test set (n = 257). As our objective was to discern the variance in model performances based on the dimensionality of features, we created three model dimensionality categories (LDM, MDM, HDM), as highlighted in Table 1.

For Low-Dimensional-Models (LDM), we solely incorporated the *(1) number of app logins per day.* This feature selection was inspired by the work of Bricker et al. (2023), who reported impressive churn prediction outcomes in two mHealth apps on the seventh day using only this particular feature [7].

Medium-Dimensional-Models (MDMs) had the following additional features available: *(2) minutes spent in the app per day, (3) number of unique app events per day, (4) number of total app events per day, (5) weekday of install, (6) user weight*, and *(7) user acquisition type.* The choice of features from the second to the fourth was to provide a broader perspective on user engagement. The selection of the fifth to the seventh features was anchored in prior works linking these features to variations in user churn [5, 27, 29, 74, 87].

Lastly, High-Dimensional-Models (HDMs) were the most comprehensive. They encompassed all the features from the LDM and MDM categories and further incorporated features detailing daily *(8) minutes spent in each app event per day.* This addition led to the inclusion of 174 more features for each day. In comparison to other engagement features, these features also include information on the specific app components users interact with *(e.g., Event 1 - User views his activity plan, Event 7 - User views game information).*

The following ML algorithms were applied to predict churn for all three model dimensionalities (LDM, MDM, HDM) on each of the first seven days of app usage: LR, DT, SVM, RF, XGB, NN, BLR and ENS that averaged LR, XGB, RF, and NN model predictions. For all models, we applied stratified 10-fold cross-validation and randomized search for hyperparameter tuning on the training set, optimizing for F1 score. Aligned with previous research, we applied Tomek Links undersampling to ensure an even distribution of churned and retained users in each fold [22]. Detailed hyperparameter grids are provided in the supplementary material. All models inherently performed feature selection (e.g., through regularization), ranking features based on their importance in making predictions. We evaluated models using F1 score, area under the curve (AUC), accuracy, precision, and recall. Finally, we compare the confusion matrix of our best-performing model from Day 1 until Day 6 with the number of users that reengaged with the app post-churn prediction within the trial phase to estimate churn intervention potential. We used freely available Python packages listed in supplementary material for data manipulation, analysis, visualization, modeling, and evaluation.

5 RESULTS

5.1 Descriptive results

Among all included users, 35% retained and 65% churned. The dataset included a total of 862 iOS users and 409 Android users with an additional 12 users accessing the app via both platforms. Android users exhibited a slightly higher conversion rate (38.1%, 156/409) compared to iOS users (33.5%, 289/862), however, this difference was not statistically significant ($\chi^2(1) = 2.40$, $p = 0.12$).

Table 2: Descriptive statistics of categorical features, retention rates, and smartphone operating systems (n = 1,283)

		All users (n = 1,283)		Retained users (n = 449)		Churned users (n = 834)	
Feature	**Category**	**n**	**%**	**n**	**%**	**n**	**%**
(5) Weekday of Install	Monday	193	15.0%	65	14.5%	128	15.4%
	Tuesday	176	13.7%	60	13.4%	116	13.9%
	Wednesday	205	16.0%	77	17.2%	128	15.4%
	Thursday	204	15.9%	73	16.3%	131	15.7%
	Friday	168	13.1%	60	13.4%	108	13.0%
	Saturday	166	12.9%	61	13.6%	105	12.6%
	Sunday	171	13.3%	53	11.8%	118	14.2%
(7) User Acquisition Type	Paid	755	58.8%	273	60.8%	482	57.8%
	Organic	528	41.2%	176	39.2%	352	42.2%
Daily active users	Day 1	1283	100.0%	449	100.0%	834	100.0%
	Day 2	435	33.9%	239	53.2%	160	23.5%
	Day 3	405	31.6%	245	54.6%	160	19.2%
	Day 4	377	29.4%	236	52.6%	141	16.9%
	Day 5	365	28.5%	234	47.9%	131	15.7%
	Day 6	329	25.6%	223	49.7%	106	12.7%
	Day 7	359	28.0%	246	54.8%	113	13.6%
Smartphone operating system	iOS	862	67.2%	289	64.4%	573	68.7%
	Android	409	31.9%	156	34.7%	253	30.3%
	Both	12	0.9%	4	0.9%	8	1.0%

Table 2 presents the descriptive statistics of categorical features and retention rates. Users who discovered the app through paid social media advertising consisted of relatively fewer churned users (63.8%, 482/755) than users who discovered the app organically (66.7%, 352/528). The highest churn rate was observed among users who began using the app on Sundays (69.0%, 118/171), while the lowest was for those who started on Saturdays (63.4%, 105/166).

Descriptive statistics of numerical features are provided in Table 3. Churned users demonstrated a lower mean *(1) number of app logins per day, (2) minutes spent in the app per day, (3) number of unique app event types per day*, and *(4) number of total app events per day* on each of the seven trial days. Additionally, churned users had a higher average *(6) user weight*, 238.12 lbs, compared to the 229.79 lbs average of retained users. A notable trend was the sharp decline in engagement from the first to the second day, followed by a more gradual decrease over the trial's seven days, with a slight uptick between Days 6 and 7. This pattern is further illustrated in the retention chart in Figure 1. Descriptive statistics of categorical and numerical features in the training and test set are provided in the supplementary material.

5.2 Model results

The performance metrics of each model, including F1 score, AUC, Accuracy, Precision, and Recall, are consolidated in Table 4. A consistent trend emerged with F1, AUC, and Accuracy improving for each model as the trial days progressed, as visualized in Figure 2. When comparing the three model dimensionalities (LDMs, MDMs, HDMs), it was evident that as more features became available, recall decreased while precision increased. High recall refers to the model's ability to identify as many churned users as possible (true positives), while high precision indicates the model's ability to minimize false alarms (false positives). The F1 score balances both and is a common metric for overall model assessment. While the differences between the various ML algorithms were marginal, LDMs stood out with an average F1 score of 0.834 over the seven days. This was slightly better than HDMs at 0.827 and MDMs at 0.823. On the first day, however, HDMs (mean F1 = 0.806) and MDMs (mean F1 = 0.803) marginally outperformed LDMs (mean F1 = 0.801).

The RF model, which only used the *(1) number of app logins per day* feature (RF-LDM) achieved the best mean F1 score across all seven days, registering at 0.839. The best-performing models on each individual day were LR-HDM (Day 1, F1 = 0.824), RF-LDM (Day 2, F1 = 0.824), SVM-MDM (Day 3, F1 = 0.843), BLR-LDM (Day 4, F1 = 0.852), LR-LDM (Day 5, F1 = 0.854), SVM-LDM (Day 6, F1 = 0.862), and RF-HDM (Day 1, F1 = 0.866), averaging to a mean F1 score of 0.846, which is only a marginal improvement compared to the performance of RF-LDM.

Higher-dimensional models were generally better at distinguishing between positive and negative classes across a range of decision thresholds, leading to a higher AUC. HDMs ranked the highest with a mean AUC of 0.789 compared to MDMs (mean AUC = 0.764) and LDMs (mean AUC = 0.705). Notably, higher dimensional models performed better at thresholds with low false positive rates. This is further highlighted in the supplementary material, where we report the Receiver Operating Characteristic (ROC) curves of all models on each day.

Table 3: Descriptive statistics of numerical features (n = 1,283)

Feature	Day	All users (n = 1,283)			Retained users (n = 449)			Churned users (n = 834)		
		Mean (SD)	Median (IQR)	Range	Mean (SD)	Median (IQR)	Range	Mean (SD)	Median (IQR)	Range
(1) Number of app logins per day	Day 1	2.7(1.7)	2(1-3)	1-11	3.3 (2.0)	3(2-4)	1-10	2.4(1.4)	2(1-3)	1-11
	Day 2	0.8(1.4)	0(0-1)	0-11	1.4 (1.9)	1(0-2)	0-11	0.4 (0.9)	0(0-0)	0-7
	Day 3	0.7(1.2)	0(0-1)	0-9	1.3(1.7)	1(0-2)	0-9	0.3 (0.8)	0(0-0)	0-7
	Day 4	0.6(1.3)	0(0-1)	0-10	1.3(1.7)	1(0-2)	0-10	0.3(0.8)	0(0-0)	0-8
	Day 5	0.6(1.2)	0(0-1)	0-10	1.3(1.7)	1(0-2)	0-10	0.2(0.6)	0(0-0)	0-5
	Day 6	0.5(1.2)	0(0-1)	0-10	1.2(1.7)	0(0-2)	0-10	0.2(0.6)	0(0-0)	0-7
	Day 7	0.6(1.3)	0(0-1)	0-9	1.3(1.7)	1(0-2)	0-9	0.2(0.7)	0(0-0)	0-7
(2) Minutes spent in the app per day	Day 1	26.7(32.5)	15.7(7.3-33.8)	0.5-279.9	38.1(44.7)	22.7(9.5-49.9)	0.8-279.9	20.6(21.2)	12.9(6.9-28.2)	0.48-154.1
	Day 2	5.3(15.0)	0(0-2.2)	0-216.4	10.9(21.3)	0.6(0-12.2)	0-216.4	2.3(8.8)	0(0-0)	0-96.2
	Day 3	4.1(13.3)	0(0-1.4)	0-189.9	8.9(19.7)	1.0(0-7.0)	0-189.9	1.5(6.6)	0(0-0)	0-99.0
	Day 4	4.1(13.5)	0(0-1.0)	0-176.3	8.7(19.6)	0.6(0-7)	0-176.3	1.6(7.5)	0(0-0)	0-126.6
	Day 5	4.0(15.2)	0(0-0.7)	0-311.3	9.3(23.4)	0.4(0-7.4)	0-311.3	1.2(6.2)	0(0-0)	0-105.5
	Day 6	3.6(14.3)	0(0-0.2)	0-225.5	9.2(22.7)	0(0-6.4)	0-225.5	0.7(3.6)	0(0-0)	0-54.1
	Day 7	3.8(12.7)	0(0-0.7)	0-136.8	9.1(19.3)	0.63(0-7.1)	0-136.8	0.9(4.8)	0(0-0)	0-71.5
(3) Number of unique app event types per day	Day 1	17.9(6.9)	17(13-21)	1-47	16.7(6.2)	16.2(12-19.8)	1-45	18.5(7.2)	17.2(13.5-2)	5-47
	Day 2	4.2(6.5)	0(0-9.7)	0-30	6.6(7.1)	6(0-12)	0-30	3.0(5.8)	0(0-0)	0-27
	Day 3	3.8(6.2)	0(0-8.5)	0-35	6.5(6.6)	7(0-12)	0-32	2.4(5.5)	0(0-0)	0-35
	Day 4	3.5(5.9)	0(0-8.3)	0-45	6.2(6.5)	6.3(0-12)	0-29	2.1(5.1)	0(0-0)	0-45
	Day 5	3.5(6.0)	0(0-8.0)	0-33	6.3(6.66)	6(0-12)	0-25	2.0(5.0)	0(0-0)	0-33
	Day 6	3.2(5.8)	0(0-5.2)	0-38	5.8(6.4)	0(0-11.5)	0-26	1.8(5.0)	0(0-0)	0-38
	Day 7	3.4(6.0)	0(0-7)	0-35	6.5(6.8)	7(0-11.5)	0-35	1.7(4.7)	0(0-0)	0-29
(4) Number of total app events per day	Day 1	133.7(114.7)	109(59-174.5)	2-1188	172.0(152.6)	135(59-245)	2-1188	113.0(80.7)	98(59-147)	6-794
	Day 2	21.2(45.2)	0(0-24.0)	0-525	39.8(61.2)	12(0-57)	0-525	11.2(29.1)	0(0-0)	0-295
	Day 3	17.8(39.3)	0(0-20.0)	0-329	35.5(52.2)	14(0-53)	0-329	8.2(25.6)	0(0-0)	0-292
	Day 4	16.2(36.6)	0(0-19.0)	0-411	32.2(47.5)	12(0-50)	0-316	7.6(25.2)	0(0-0)	0-411
	Day 5	15.9(36.4)	0(0-14.0)	0-327	33.5(50.0)	9(0-56)	0-327	6.4(20.8)	0(0-0)	0-220
	Day 6	14.1(37.4)	0(0-7.0)	0-540	30.6(54.6)	0(0-47)	0-540	5.2(18.1)	0(0-0)	0-178
	Day 7	16.1(41.6)	0(0-12.5)	0-658	36.1(61.2)	13(0-54)	0-658	5.2(17.8)	0(0-0)	0-232
(6) User weight (lbs)	-	235.2(62.3)	229(190-270)	97-536	229.8(58.9)	225(188-260)	114-477	238.1(63.9)	230(194-275)	97-536

SD = standard deviation; IQR = interquartile range

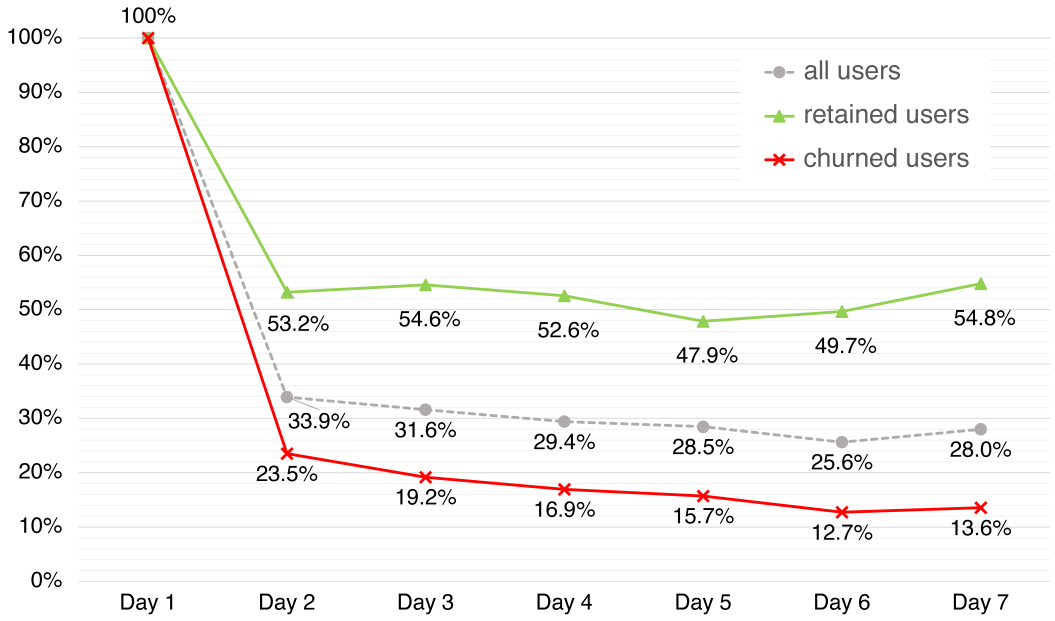

Figure 1: Retention chart, displaying the percentages of users who log in on each day of the trial period (n = 1,283)

Table 4: Churn prediction model results on the test set for each day of the 7-day trial phase (n = 257).

		Low-Dimensional Model (LDM)									Medium-Dimensional Model (MDM)									High-Dimensional Model (HDM)								
Metric	**Day**	**LR**	**DT**	**SVM**	**RF**	**XGB**	**NN**	**ENS**	**BLR**	***mean***	**LR**	**DT**	**SVM**	**RF**	**XGB**	**NN**	**ENS**	**BLR**	***mean***	**LR**	**DT**	**SVM**	**RF**	**XGB**	**NN**	**ENS**	**BLR**	***mean***
F1 score	Day 1	0.801	0.801	0.809	0.801	0.801	0.780	0.809	0.809	*0.801*	0.789	0.806	0.808	0.811	0.798	0.802	0.807	0.803	*0.803*	**0.824**	0.820	0.818	0.802	0.791	0.786	0.815	0.789	*0.806*
	Day 2	0.814	0.813	0.803	**0.824**	0.818	0.814	0.803	0.814	*0.813*	0.803	0.760	0.806	0.773	0.793	0.754	0.801	0.808	*0.787*	0.807	0.813	0.822	0.817	0.798	0.795	0.812	0.798	*0.808*
	Day 3	0.827	0.839	0.834	0.840	0.839	0.822	0.837	0.827	*0.833*	0.820	0.780	**0.843**	0.832	0.816	0.823	0.835	0.811	*0.820*	0.824	0.808	0.832	0.835	0.812	0.819	0.829	0.815	*0.822*
	Day 4	0.850	0.819	0.849	0.840	0.832	0.801	0.842	**0.852**	*0.836*	0.826	0.839	0.828	0.831	0.801	0.789	0.829	0.824	*0.821*	0.810	0.813	0.827	0.834	0.820	0.830	0.822	0.830	*0.823*
	Day 5	**0.854**	0.848	0.833	0.852	0.850	0.828	0.847	0.853	*0.846*	0.838	0.848	0.848	0.837	0.830	0.837	0.853	0.832	*0.840*	0.818	0.827	0.841	0.853	0.829	0.821	0.832	0.833	*0.832*
	Day 6	0.859	0.852	**0.862**	0.848	0.845	0.844	0.853	0.859	*0.853*	0.856	0.853	0.851	0.847	0.832	0.807	0.857	0.856	*0.845*	0.835	0.827	0.850	0.846	0.843	0.856	0.853	0.844	*0.844*
	Day 7	0.854	0.853	0.857	0.865	0.865	0.852	0.857	0.839	*0.855*	0.833	0.848	0.847	0.860	0.827	0.848	0.853	0.834	*0.844*	0.838	0.853	0.857	**0.866**	0.854	0.846	0.862	0.844	*0.852*
	mean	*0.837*	*0.832*	*0.835*	***0.839***	*0.836*	*0.820*	*0.835*	*0.836*	*0.834*	*0.824*	*0.819*	*0.833*	*0.827*	*0.814*	*0.809*	*0.834*	*0.824*	*0.823*	*0.822*	*0.823*	*0.835*	*0.836*	*0.821*	*0.822*	*0.832*	*0.822*	*0.827*
AUC	Day 1	0.677	0.655	0.571	0.669	0.659	0.677	0.677	0.677	*0.658*	0.706	0.696	0.679	0.706	0.733	0.738	0.733	0.693	*0.711*	0.746	0.656	0.776	0.766	0.770	0.780	**0.782**	0.759	*0.755*
	Day 2	0.678	0.659	0.678	0.678	0.675	0.684	0.682	0.678	*0.677*	0.727	0.713	0.698	0.754	0.747	0.690	0.750	0.734	*0.727*	0.750	0.753	0.746	0.776	**0.788**	0.742	0.773	0.767	*0.762*
	Day 3	0.704	0.719	0.704	0.696	0.698	0.664	0.702	0.708	*0.699*	0.755	0.748	0.785	0.780	0.782	0.684	0.777	0.751	*0.758*	0.776	0.796	0.792	0.769	**0.801**	0.771	0.794	0.777	*0.785*
	Day 4	0.723	0.683	0.719	0.713	0.707	0.695	0.708	0.729	*0.710*	0.765	0.739	0.774	0.799	0.779	0.745	0.785	0.766	*0.769*	0.793	0.765	0.798	0.713	**0.821**	0.811	0.816	0.784	*0.788*
	Day 5	0.736	0.711	0.704	0.726	0.716	0.716	0.732	0.736	*0.722*	0.781	0.769	0.800	0.808	0.801	0.704	0.812	0.786	*0.783*	0.802	0.704	0.806	0.743	0.812	0.832	**0.837**	0.819	*0.794*
	Day 6	0.749	0.733	0.746	0.731	0.724	0.730	0.735	0.747	*0.737*	0.797	0.775	0.816	0.822	0.805	0.761	0.813	0.806	*0.799*	0.813	0.705	0.823	0.829	0.859	0.841	**0.860**	0.843	*0.821*
	Day 7	0.735	0.735	0.742	0.731	0.727	0.737	0.729	0.742	*0.735*	0.785	0.784	0.792	0.800	0.821	0.810	0.805	0.799	*0.799*	0.789	0.773	0.800	0.749	0.856	**0.864**	0.856	0.855	*0.818*
	mean	*0.715*	*0.699*	*0.695*	*0.706*	*0.701*	*0.700*	*0.709*	*0.717*	*0.705*	*0.759*	*0.746*	*0.763*	*0.781*	*0.781*	*0.733*	*0.782*	*0.762*	*0.764*	*0.781*	*0.736*	*0.792*	*0.764*	*0.815*	*0.806*	***0.817***	*0.801*	*0.789*
Accuracy	Day 1	0.685	0.685	0.708	0.685	0.685	0.685	0.685	0.708	*0.691*	0.696	0.708	0.716	0.716	0.708	0.720	0.716	0.716	*0.712*	**0.743**	0.724	0.739	0.704	0.720	0.716	**0.743**	0.708	*0.725*
	Day 2	0.720	0.720	0.697	0.716	0.724	0.720	0.720	0.720	*0.717*	0.708	0.681	0.700	0.700	0.716	0.677	0.720	0.716	*0.702*	0.720	0.720	**0.743**	0.735	0.728	0.724	0.739	0.720	*0.729*
	Day 3	0.739	0.759	0.747	0.763	0.759	0.759	0.747	0.739	*0.751*	0.739	0.712	**0.767**	0.759	0.743	0.739	0.755	0.724	*0.742*	0.751	0.747	0.755	0.759	0.751	0.751	0.755	0.747	*0.752*
	Day 4	0.778	0.739	0.778	0.759	0.759	0.767	0.747	**0.782**	*0.764*	0.747	0.759	0.751	0.755	0.724	0.739	0.755	0.743	*0.747*	0.728	0.743	0.747	0.755	0.763	0.759	0.751	0.767	*0.751*
	Day 5	0.786	0.778	0.759	0.782	0.782	0.782	0.720	0.786	*0.772*	0.770	0.770	0.786	0.767	0.774	0.767	**0.790**	0.763	*0.773*	0.747	0.755	0.770	0.786	0.770	0.767	0.767	0.774	*0.767*
	Day 6	0.798	0.786	**0.802**	0.790	0.778	0.790	0.763	0.798	*0.788*	0.794	0.782	0.786	0.778	0.774	0.732	0.794	0.794	*0.779*	0.770	0.755	0.782	0.778	0.782	0.794	0.790	0.790	*0.780*
	Day 7	0.790	0.790	0.794	**0.805**	**0.805**	0.790	0.778	0.770	*0.790*	0.767	0.790	0.782	0.798	0.770	0.782	0.790	0.767	*0.781*	0.770	0.790	0.794	**0.805**	0.802	0.790	**0.805**	0.790	*0.793*
	mean	*0.757*	*0.751*	*0.755*	*0.757*	*0.756*	*0.756*	*0.737*	*0.758*	*0.753*	*0.746*	*0.743*	*0.755*	*0.753*	*0.744*	*0.736*	*0.760*	*0.746*	*0.748*	*0.747*	*0.748*	*0.762*	*0.760*	*0.759*	*0.757*	***0.764***	*0.757*	*0.757*
Precision	Day 1	0.694	0.694	0.720	0.694	0.694	0.749	0.720	0.720	*0.710*	0.737	0.726	0.737	0.730	0.744	0.760	0.739	0.749	*0.740*	0.762	0.727	0.767	0.726	0.791	**0.793**	0.788	0.765	*0.765*
	Day 2	0.732	0.736	0.710	0.733	0.731	0.703	0.710	0.732	*0.723*	0.732	0.765	0.711	0.784	0.773	0.770	0.763	0.737	*0.754*	0.750	0.736	0.768	0.760	**0.793**	0.789	0.784	0.772	*0.769*
	Day 3	0.744	0.757	0.744	0.755	0.759	0.709	0.746	0.744	*0.745*	0.761	0.799	0.767	0.781	0.785	0.754	0.761	0.749	*0.769*	0.781	0.820	0.769	0.770	**0.821**	0.797	0.777	0.799	*0.792*
	Day 4	0.775	0.764	0.781	0.775	0.778	0.783	0.781	0.782	*0.777*	0.766	0.757	0.770	0.786	**0.834**	0.786	0.777	0.760	*0.779*	0.760	0.794	0.764	0.763	0.832	0.787	0.787	0.811	*0.787*
	Day 5	0.785	0.783	0.775	0.794	0.790	0.780	0.786	0.791	*0.786*	0.793	0.763	0.810	0.786	0.835	0.786	0.801	0.795	*0.796*	0.789	0.782	0.784	0.791	0.827	**0.842**	0.804	0.824	*0.805*
	Day 6	0.806	0.794	0.807	0.796	0.795	0.798	0.801	0.806	*0.800*	0.802	0.779	0.797	0.786	0.828	0.778	0.799	0.802	*0.796*	0.805	0.782	0.790	0.792	0.815	0.805	0.821	**0.839**	*0.806*
	Day 7	0.798	0.801	0.799	0.808	0.808	0.794	0.799	0.790	*0.800*	0.798	0.801	0.799	0.800	0.834	0.796	0.801	0.795	*0.803*	0.796	0.801	0.799	0.805	**0.842**	0.832	0.821	0.839	*0.817*
	mean	*0.762*	*0.761*	*0.762*	*0.765*	*0.765*	*0.759*	*0.763*	*0.766*	*0.763*	*0.770*	*0.770*	*0.770*	*0.779*	*0.805*	*0.776*	*0.777*	*0.769*	*0.777*	*0.778*	*0.778*	*0.777*	*0.772*	***0.817***	*0.806*	*0.797*	*0.807*	*0.792*
Recall	Day 1	0.948	0.948	0.924	0.948	0.948	**0.965**	0.924	0.924	*0.941*	0.849	0.907	0.895	0.913	0.860	0.849	0.890	0.866	*0.879*	0.895	0.942	0.878	0.895	0.791	0.779	0.843	0.814	*0.855*
	Day 2	0.919	0.907	0.924	**0.942**	0.930	0.895	0.924	0.919	*0.920*	0.890	0.756	0.930	0.762	0.814	0.738	0.843	0.895	*0.828*	0.872	0.907	0.884	0.884	0.802	0.802	0.843	0.826	*0.852*
	Day 3	0.930	0.942	0.948	0.948	0.936	**0.977**	0.954	0.930	*0.945*	0.890	0.762	0.936	0.890	0.849	0.907	0.924	0.884	*0.880*	0.872	0.797	0.907	0.913	0.802	0.843	0.890	0.831	*0.857*
	Day 4	**0.942**	0.884	0.930	0.919	0.895	0.820	0.913	0.936	*0.905*	0.895	0.942	0.895	0.901	0.831	0.785	0.890	0.901	*0.880*	0.866	0.831	0.901	0.919	0.808	0.878	0.861	0.849	*0.864*
	Day 5	0.936	0.924	0.901	0.919	0.919	0.884	0.919	0.924	*0.916*	0.890	**0.953**	0.890	0.895	0.820	0.895	0.913	0.878	*0.892*	0.849	0.878	0.907	0.924	0.831	0.802	0.861	0.843	*0.862*
	Day 6	0.919	0.919	0.924	0.907	0.901	0.895	0.913	0.919	*0.912*	0.919	**0.942**	0.913	0.919	0.837	0.837	0.924	0.919	*0.901*	0.866	0.878	0.919	0.907	0.872	0.913	0.913	0.849	*0.890*
	Day 7	0.919	0.913	0.924	0.930	0.930	0.919	0.924	0.895	*0.919*	0.872	**0.942**	0.901	0.930	0.820	0.907	0.913	0.878	*0.895*	0.884	0.913	0.924	0.936	0.866	0.861	0.907	0.849	*0.892*
	mean	*0.930*	*0.919*	*0.925*	***0.930***	*0.923*	*0.908*	*0.924*	*0.921*	*0.923*	*0.886*	*0.886*	*0.909*	*0.887*	*0.833*	*0.845*	*0.899*	*0.889*	*0.879*	*0.872*	*0.878*	*0.903*	*0.911*	*0.825*	*0.840*	*0.874*	*0.837*	*0.867*

AUC = Area under the curve; LR = Logistic Regression; DT = Decision tree; SVM = Support Vector Machine; RF = Random Forest; XGB = XGBoost; NN = Neural Network; ENS = Ensemble method; BLR = Bayesian Logistic Regression

5.3 Feature importance

Across all models, user engagement-related features were consistently considered most important, particularly when they referred to later days in the 7-day trial period. For instance, in the RF-LDM, *(1) number of app logins per day* were rated as more important and thus more predictive if the feature referred to a later day in the 7-day trial period. To illustrate, the Day-7-RF-LDM had the following feature importance's for *(1) number of app logins per day*: Day 7 (31.1%), Day 5 (20.3%), Day 4 (14.7%), Day 3 (12.3%), Day 6 (12.1%), Day 2 (5.4%), Day 1 (4.2%). *(1) Number of app logins per day* were also among the 10 most important feature in the RF-MDM except for Day 6. The RF-HDM also prioritized *(1) number of app logins per day*; however, the combined importance of this feature did not exceed 5% on any day of the RF-HDM.

The other constructed engagement features, *(2) minutes spent in the app per day*, *(3) number of unique app events per day*, *(4) number of total app events per day*, were also selected by RF-LDMs and RF-HDM. For these features, the same trend of prioritizing more recent features, particularly from the preceding two days before the day of prediction, could be observed. The features *(5) weekday of install, (6) user weight*, and *(7) user acquisition type* only had a minor importance in MDMs on the first two days (combined importance Day 1 = 25.6%, Day 2 = 15.9%). From Day 3 forward, these features only accounted for a combined importance of less than 5 percent in the RF-MDM with the importance further declining for each consecutive day. In the RF-HDM, *(5) weekday of install, (6) user weight*, and *(7) user acquisition type* were only selected in the first two days with combined relative importance of less than 1 percent.

The RF-HDM also prioritized several events from feature *(8) minutes spent in each app event per day*, especially those tied to overall app engagement and game activity events (e.g. ordered by importance: *Event 7 - User views game information, Event 19 - In-game submission caption screen, Event 1 - User views his activity plan, Event 4 - User opens the app, Event 20 - In game submission confirmation screen, Event 5 - User backgrounds the app, Event 17 - In game submission preview screen*). Despite the detailed tracking of 174 different app events, it's noteworthy that these granular app engagement features did not enhance our models' capability in detecting churned users (true positives) and likewise did not increase F1 score, except marginally on the first day. However, the

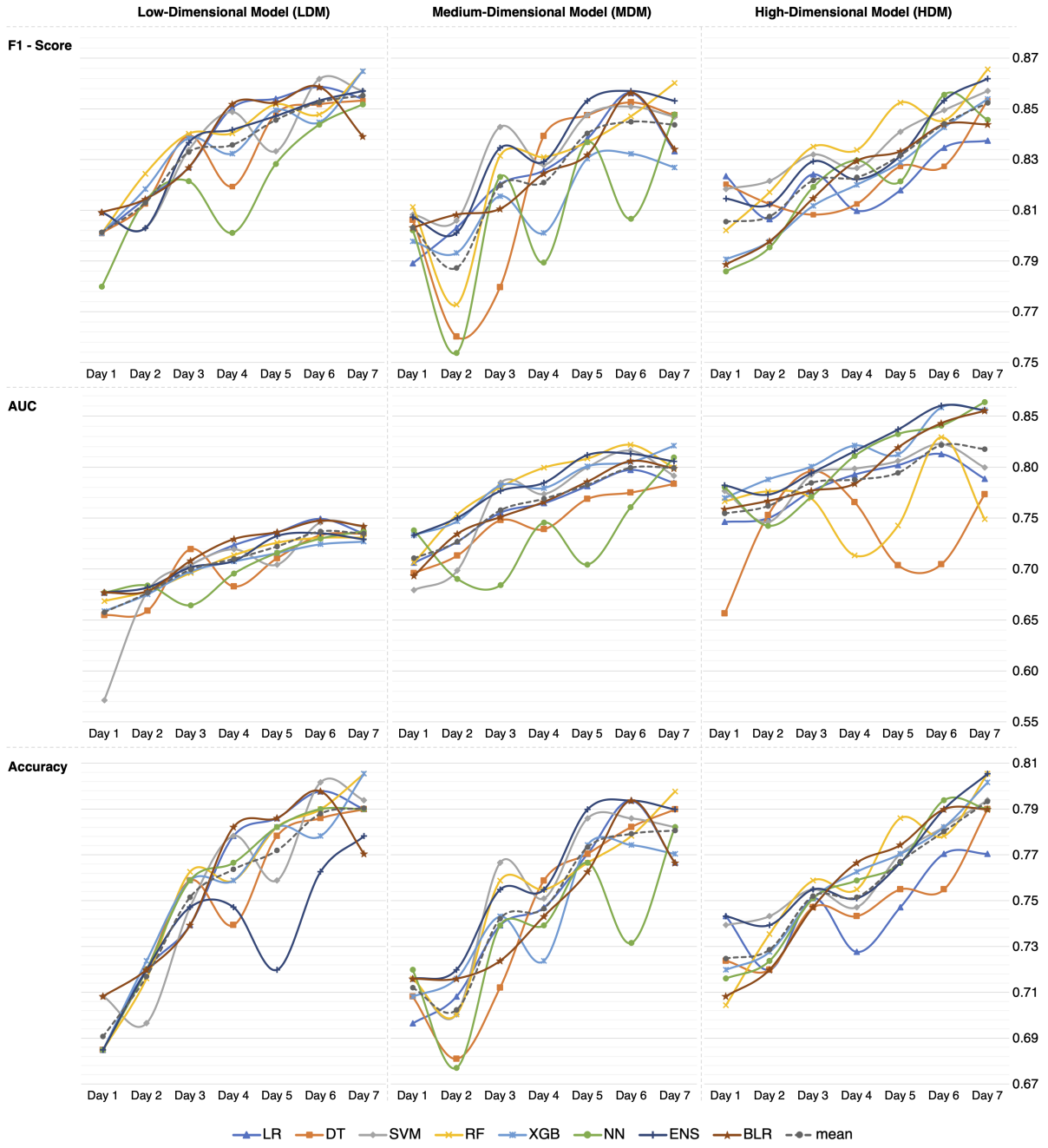

Figure 2: F1 score, AUC, and Accuracy of applied ML algorithms for churn prediction on each day of the 7-day trial phase in the test set (n = 257).

application of these features resulted in models' improved ability to differentiate between the positive and negative classes across various decision thresholds, particularly at thresholds with low false positive rates, thus improving AUC. In the supplementary material we provide a detailed overview of the most important features in the RF-LDM, RF-MDM, and RF-HDM on each day of the seven trial days.

5.4 Intervention potential

For the RF-LDM, we compare the prediction confusion matrix on Days 1-6 with user activity on subsequent days up to the trial's end on Day 7, as showcased in Table 5. As depicted in Figure 3, the proportion of churned users who remained active in the app post-churn-prediction (reengaged churned users) consistently decreased

Table 5: Confusion matrix of RF-LDM churn prediction on Days 1-7 on the test set, compared with user activity on consecutive days (n = 257)

User reengagement after prediction		Prediction on previous day (Day 1-6)					
Time window	Reengagement	True Positive	True Negative	False Positive	False Negative	n	%
Day 2-7	TRUE	76	13	49	8	146	57%
	FALSE	87	0	23	1	111	43%
Day 3-7	TRUE	66	25	36	10	137	53%
	FALSE	96	1	23	0	120	47%
Day 4-7	TRUE	59	32	27	8	126	49%
	FALSE	104	0	26	1	131	51%
Day 5-7	TRUE	43	38	19	10	110	43%
	FALSE	115	1	27	4	147	57%
Day 6-7	TRUE	30	44	11	12	97	38%
	FALSE	128	0	30	2	160	62%
Day 7	TRUE	21	43	7	11	82	32%
	FALSE	135	2	33	5	175	68%
Prediction at the end of Day 7		160	47	38	12	257	100%

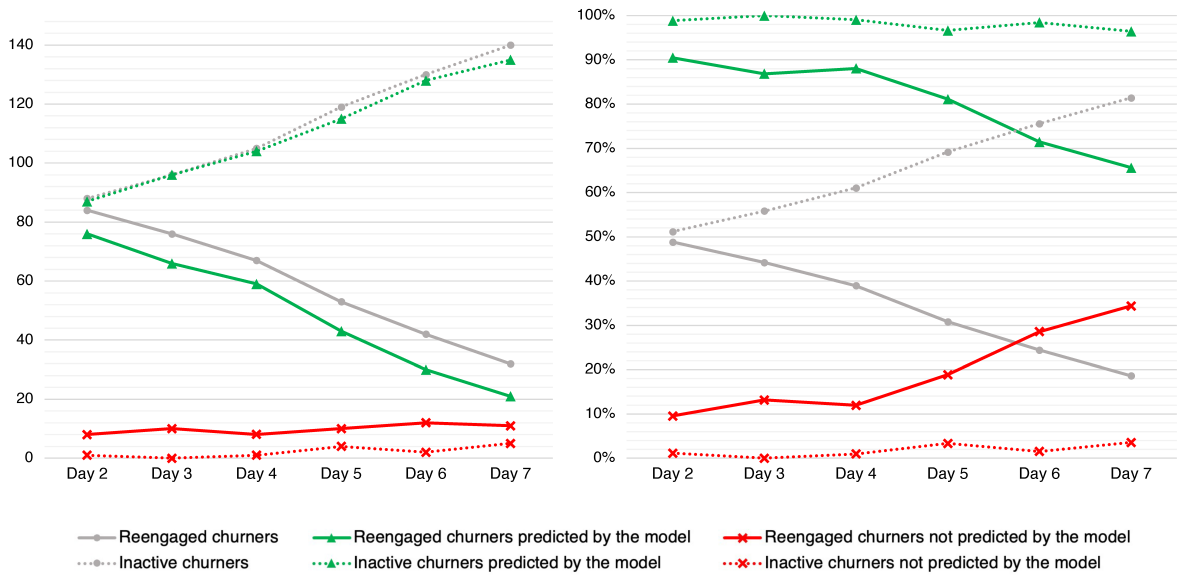

Figure 3: Number of churned users (left-side) and percentage of churned users (right-side) in the test set who reengaged with the app post-churn-prediction (n = 172)

from Day 2 (48.8%) to Day 7 (18.6%). Conversely, the number of churned users who didn't return to the app post-churn-prediction (inactive churned users) increased from Day 2 (51.2%) to Day 7 (81.4%). The RF-LDM was especially adept at identifying inactive churned users, with a mean accuracy of 98.2% across the seven days. The model correctly predicted a smaller and overall declining percentage of reengaged churned users from Day 2 (90.5%) to Day 7 (65.6%). Across the first 6 days, the RF-LDM correctly predicted 80.6% of reengaged churned users.

Overall, the model correctly predicted an average of 93% of churned users over the trial week. However, it is worth noting that while the RF-LDM was effective in predicting churned users, it also had a significant number of false positives, with the percentage of retained users misclassified as churned users decreasing from 85% on Day 1 to 45% on Day 7, averaging 58.7% over the trial week.

6 DISCUSSION

6.1 Model comparison and feature importance

We assessed churn prediction models for a widely used weight loss app over the initial seven trial days using diverse ML algorithms (LR, DT, SVM, RF, XGB, NN, ENS, BLR) across three feature

dimensionality categories (LDM, HDM, HDM). The performance differences among the various ML algorithms we employed were marginal across selected feature dimensions. Hence, determining the best algorithm for churn prediction remains ambiguous. Several studies have evaluated and compared various ML algorithms for churn prediction, coming to different conclusions [15, 50, 73, 79]. Thus, the optimal algorithm likely depends on the dataset's nature and features. Developers should thus experiment with multiple algorithms, considering both model interpretability and complexity.

Consistent with prior churn prediction studies in DHIs, our models' predictive performance was predominantly driven by user app engagement data, with *(1) number of app logins per day* capturing app engagement effectively, as previously proposed by Bricker et al. (2023) [7]. Our LDMs, which focused on this metric, yielded comparable F1 scores to MDMs and HDMs that incorporated a broader range of engagement and user-related features. Adding more granular app engagement features, such as *(8) minutes spent in each app event type per day*, did not improve our models' capability in detecting churned users (true positives) and likewise did not increase the F1 score, except marginally on the first day. However, our higher-dimensional models were better suited to differentiating between the positive and negative classes across various decision thresholds, thus achieving higher AUC values. Higher dimensional models performed better at thresholds with low false positive rates. Our findings also demonstrate that churn prediction performance improves as more user engagement data becomes available over time. Our models generally improved in terms of F1, AUC, and Accuracy from Day 1 to Day 7. This progressive improvement in churn prediction over extended periods aligns with intuitive expectations and was also observed in another mHealth app [39].

The static user-related features *(5) weekday of install,* (6) *user weight, and (7) user acquisition type,* which we selected based on prior research connecting these features to mHealth app adherence, did not improve our models' performance. While our results align with previous research indicating that greater disease severity, and thus likely greater weight, is associated with lower adherence [5, 29, 74], the differences weren't significant enough to enhance our models. Previous studies also report that downloading and starting mHealth apps on weekends, as opposed to weekdays, is associated with reduced adherence [27, 29]. However, inconclusively, users who downloaded the app on Sundays had the lowest percentage of retained users, while users downloading the app on Saturdays resulted in the highest percentage of retained users in our dataset. We also observed differences in retention depending on the type of user acquisition channel. However, the differences were insufficient to improve our models [87]. Examples of features from other studies that may further enhance the performance of churn prediction models, however unavailable in our case, include vectors of user messages or user reviews [39], push notification responses [20], or, if the app is distributed personally, intervention provider information [62].

Our best-performing model in detecting churned users (true positives) was a RF model utilizing only *(1) number of app logins per day* as features (RF-LDM). This model achieved an F1 score of 0.865 (AUC = 0.731, Accuracy = 80.5%) on Day 7 and an average F1 score of 0.839 over the initial week (mean AUC = 0.706, mean Accuracy = 0.757). The model correctly identified on average 93% of churned users with a mean false positive rate of 58.7%. Given these results, our study aligns with previous research, concluding that churn prediction in mHealth apps with user app engagement data is viable. However, drawing direct comparisons between individual churn prediction studies is difficult due to variations in health interventions, churn definitions, and prediction windows used across studies. Notably, the studies by Kwon et al. (2021) and Bricker et al. (2023) bear the closest resemblance to ours [7, 39]. Kwon et al. (2021) reported an F1 Score of 0.83 (AUC = 0.82, Accuracy = 83%, excluding text vector feature) in predicting users seeking refunds in a weight loss intervention after 16 weeks. Yet, when the prediction window was shortened by the week prior to churning, the F1 score dropped to 0.68 (AUC = 0.70, Accuracy = 70%), which is lower than our Day 7 prediction. Bricker et al. (2023) achieved comparably higher churn prediction results after seven days using only daily login counts in two smoking cessation apps, "iCanQuit" (F1 = 0.896, AUC = 0.940, Accuracy = 88.5%, classification threshold = 0.5) and "QuitGuide" (F1 = 0.861, AUC = 0.880, Accuracy = 81.16%, classification threshold = 0.5), employing LR. Our approach to modeling LR-LDM differed slightly, especially in using Tomek Links sampling during cross-validation to counteract bias towards the majority class of churned users. However, this adjustment doesn't fully account for the performance disparity. Without sampling, our Day 7 LR-LDM model's performance saw only a slight improvement (F1 Score: 0.859, AUC = 0.741, Accuracy: 79.4%).

6.2 The potential of churn prediction for targeted churn interventions

As demonstrated by our and previous studies, churn prediction holds promise to guide personalized and timely churn interventions that are likely more effective than common rule-based interventions, like sending a push notification after seven days of inactivity [7, 20, 39, 54, 62, 78]. By analyzing user engagement post-churn-prediction during the trial period, we derived valuable insights for implementing churn-prediction-driven interventions. Churn predictions in the early days of the user's journey are more impactful, as a larger fraction of users at risk of churning are still active and can benefit from in-app churn interventions. However, as more data accumulates, the performance of predictions improves. Interestingly, our models were particularly adept at identifying users who would not reengage post-churn prediction. Therefore, developers should also consider external communication channels, such as emails, when implementing churn interventions.

The choice of a churn prediction model and its features also depends on the nature of the churn prevention strategy. For interventions with minimal adverse effects, like reminder push notifications, it is less problematic if retained users are falsely predicted as churned users (false positives). In this case, the focus should be on accurately detecting churned users (true positives). In our study, a RF model that only used *(1) number of app logins per day* as features was particularly well suited for this use case while also offering advantages in terms of explainability, maintenance, risk of overfitting, training times, and computational efficiency, compared to other models.

Conversely, a more conservative prediction approach is advisable to reduce false positives for churn interventions with more significant implications, like offering subscription discounts. We found that models with a richer feature set were better suited for this scenario. As the number of app engagement features increased, recall decreased while precision increased. Consequently, models with additional engagement features, such as our MDMs and HDMs, are recommended for strategies that prioritize minimizing false positives. Furthermore, our higher-dimensional models were better suited to differentiating between the positive and negative classes across various decision thresholds, thus achieving a higher AUC. This suggests that higher-dimensional models are more fitting for developers aiming to apply churn interventions at multiple thresholds, especially those with low false positive rates. However, it's crucial to note that these higher-dimensional models, given their dependency on numerous app events, necessitate a robust data pipeline for model deployment. They are likewise more susceptible to alterations in the app and data infrastructure.

Our study contributes to the limited body of work examining churn prediction in DHIs, suggesting that early user churn in DHIs can be forecasted with user app engagement data. Most importantly, our results indicate that a substantial percentage of churned users can be identified before they cease to engage with the app. Therefore, churn prediction holds the potential to facilitate timely and tailored churn interventions. However, a noticeable research gap in studies that deploy churn prediction models in tandem with churn interventions in prospective trials remains.

6.3 Implications for HCI research

The demonstrated ability of our models, to correctly detect a substantial number of churned users before they cease to engage with the intervention in the first seven days, offers an early indication that churn prediction can guide timely adaptations in persuasive and behavior-change systems. This aligns with HCI research emphasizing the need for tailoring and adapting these systems to various user profiles, contexts, and behavior change stages to enhance their effectiveness [18, 32, 34, 40, 59, 61, 64, 65, 84]. Adapting systems to better match these elements can itself be a powerful strategy against churn. Our results reinforce the potential of churn prediction in informing early system adaptations, such as modifying persuasive strategies.

While our study focused on the initial week of usage, previous churn prediction studies in mHealth apps have shown commendable churn prediction accuracy over longer periods [7, 20, 39, 54, 62, 78], also emphasizing that prediction accuracy improves over time as more data accumulates [39]. This pattern was also observed in our study during the initial seven-day user interaction phase. Therefore, it is likely that our observation — where a substantial number of users correctly predicted to churn reengaged with the app within the first seven days — also holds true over longer durations. Future studies could explore combining churn prediction with persuasive and behavior-change systems that adapt when users at risk of churning are detected. Given that this approach leads to more effective interventions, it could accelerate the shift from merely descriptive analyses of user behavior to predictive and prescriptive strategies, warranting further insights into how users respond to tailored system adaptations.

Churn prediction models could also inform the timing of active interventions, such as reminder push notifications, that are particularly important in the early stages of behavior change [45, 46] and have been linked to disengagement in JITAIs when sent in inopportune moments [61]. Additionally, incorporating churn prediction could enhance personal support complementary to DHIs. The resulting churn risk score from these prediction models could guide eHealth coaches in adjusting their support strategies more effectively, addressing the limitations of their time and resources in analyzing user data [48, 68]. Churn risk scores from prediction models could also be utilized as input data in context-aware recommender systems potentially improving their ability in tailoring health suggestions [43, 76].

Understanding the reasons behind user disengagement from persuasive and behavior-change systems remains a pivotal area in HCI research [10, 53]. The ability to detect users on the brink of early churn before complete disengagement creates opportunities for deploying targeted user questionnaires, gathering vital feedback on the causes of disengagement. If integrated in future research, this feedback can provide insights into factors influencing early churn in real-world contexts where this information is otherwise hardly attainable. Given that future studies replicate our results over extended periods, this process may also enable timely questionnaires through maintained behavior change phases that aim to differentiate between unintended and happy abandonment before users cease to engage with DHIs.

Our findings also highlights that app engagement features are more indicative of churn than static features like weight, acquisition channels, or the weekday of install, previously reported to influence mHealth app adherence [5, 27, 29, 74, 87]. This emphasizes the need for future research to extend the focus on engagement patterns for behavior prediction and system adaptations, rather than solely on static user information. It is essential to recognize that churn prediction models, particularly those that demonstrate high performance in mHealth apps, are dependent on rich user app engagement data. This dependency is corroborated by our study and others in the field [7, 20, 39, 54, 62, 78]. Therefore, these models are most suitable for systems designed for regular, ongoing use, where user engagement data is naturally and continuously generated. In environments where such data is less readily available or less rich, the effectiveness of churn prediction models might be diminished.

In summary, this study introduces churn prediction as a promising tool for transforming user engagement data into actionable insights in DHIs, potentially supporting the shift from "one-size-fits-all" solutions to more personalized, context-sensitive, and adaptive persuasive and behavior-change systems.

7 LIMITATIONS AND FUTURE WORK

Our retrospective study relied on historical mobile app user data to discern churn signals and predict user churn. Prospective trials are essential to validate the real-world applicability of these models. While our findings align with prior churn prediction studies in

mHealth and other app domains, they are not universally applicable. Factors like the definition of churn, prediction time window, and available features can influence outcomes and vary by app. Specifically, our churn definition of unsubscribing from the app during a trial period does not transfer to freemium apps that do not employ a subscription model. In our context, unsubscribing also implied users avoiding a fee for a six-month subscription, linking churn to financial considerations, which is a non-factor in free apps. Offering monetary wagers on achieving specific behavioral health goals as an intervention component is another factor influencing users' decision to retain or churn which is not a common component in other health apps. However, our churn prediction results were comparable with other studies in free apps [7, 39], suggesting potential applicability in other app environments. Further research in diverse settings is necessary to substantiate the generalizability of our results. Our study's limited scope and the recent nature of our dataset also restricted our ability to assess the long-term performance of our churn prediction models, highlighting a potential avenue for future research. Particularly, there is a need for studies evaluating extended churn prediction periods that analyze user reengagement following accurate churn predictions.

Two primary limitations might explain our study's lower churn prediction performance compared to Bricker et al (2023) [7]. Firstly, we noticed that some users demonstrated unusual behavior, which interfered with prediction performance. Specifically, 26% (117/449) of retained users did not log into the app between Day 2 and Day 7, refraining from engaging with the app during the trial period but converting to paid subscribers. This might be due to these users not being available in the trial period but rather committing to use the app in the future or forgetting to unsubscribe and not requesting a refund. When excluding these 117 users, the performance of our RF-LDM model improved significantly on day 7 (F1 score = 0.89, AUC = 0.87, Accuracy 0.84, Precision = 0.87, Recall = 0.92), matching the results of Bricker et al (2023) [7]. Secondly, unlike the apps in the study by Bricker et al. (2023), WayBetter implements rule-based churn interventions during the trial (e.g., an email reminder sent to users who started but did not complete certain app events), which our models could not factor in. These churn interventions could have altered user engagement patterns, potentially negatively affecting the performance of our models. This also highlights a general limitation with currently applied churn prediction models: their performance will likely diminish when churn interventions are introduced unless they are accounted for in the model. Another general limitation of currently applied churn prediction methods is that significant app updates can alter user behavior, necessitating model retraining and reevaluation.

Churn prevention systems utilizing reinforcement learning may address these limitations. In this approach, the churn prediction model's risk score would inform the state of an RL agent. The RL agent then explores and exploits sequential churn interventions, with the policy updating based on user reactions to these interventions (e.g., a user login shortly after the intervention) and the intervention history, leading to more personalized and adaptive interventions [44, 72, 77]. Such advanced methodologies may incorporate Bayesian algorithms that can also obtain measures of confidence for the prediction. Our demonstration that BLR performs comparably to standard ML algorithms in churn prediction highlights a first step in this direction. Notably, there are no studies in the academic literature that have combined churn prediction in DHIs with churn interventions prospectively in a controlled study environment. Such studies, which ideally encompass a control group (no churn intervention) and a benchmark group reflecting commonly applied rule-based churn interventions (e.g., a churn intervention triggered after a prespecified period of inactivity) will be substantial in validating the real-world applicability of churn prediction models in DHIs.

8 CONCLUSION

We evaluated user churn prediction models for a public weight loss app, applying eight ML algorithms across three feature dimension sets. While the differences in performance across applied algorithms were marginal, determining an optimal algorithm for churn prediction remains challenging. The best algorithm is contingent on the specific nature of the dataset and the features incorporated. The predictive performance of our models can be attributed to app engagement features, with daily user login counts capturing app engagement effectively in our case. Our best-performing model, a Random Forest model utilizing only users' number of app logins per day as features, achieved an F1 score of 0.865 on Day 7 and an average of 0.839 over the initial week. The model correctly identified 98.2% of churned users who became inactive in the first seven days after the prediction. Notably, across the first six days, the model also captured 80.6% of churned users who remained active in the app post-churn-prediction within the first seven days. However, while the model effectively predicted churned users, it also had many false positives, averaging 58.7% over the trial week. Adding more granular app engagement features, such as users' minutes spent in each of 174 app events per day, did not improve our models' capability in detecting churned users (true positives) and likewise did not increase the F1 score, except marginally on the first day. However, our higher-dimensional models were better suited at differentiating between the positive and negative classes across various decision thresholds, thus achieving higher AUC values. This suggests that higher-dimensional models are more fitting for developers aiming to apply churn interventions at multiple thresholds, especially those with low false positive rates. When applying churn prediction models to inform personalized churn interventions, developers need to carefully consider the adverse effects of churn interventions and finetune their models accordingly. Additional static features we integrated — drawn from prior research connecting them to app adherence, namely the weekday of installation, user weight, and acquisition type — did not improve model performance. Comparing our results with similar studies, we observed performance disparities that may be attributed to unique user behaviors and rule-based churn interventions in the analyzed app. If not accounted for, such interventions can skew the performance of churn prediction models. This highlights the need for future models to be more adaptive. In conclusion, our results indicate that churn prediction in DHIs holds great potential in guiding timely, personalized interventions, enhancing user retention, and ultimately health outcomes. More prospective studies are needed to validate the real-world applicability of these models for the prevention of user churn.

Conflicts of Interest

All authors are affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich; the Department of Management, Technology, and Economics at Swiss Federal Institute of Technology Zurich; and the Institute of Technology Management and School of Medicine at the University of St. Gallen. The Centre for Digital Health Interventions is funded in part by CSS, a Swiss health insurer; Uniqa, an Austrian health insurer; and MTIP, a Swiss digital health investor. EF and TK are also co-founders of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, neither CSS, Uniqa, MTIP, or Pathmate Technologies were involved in this research.

REFERENCES

[1] Ammara Ahmed and D. Maheswari Linen. 2017. A review and analysis of churn prediction methods for customer retention in telecom industries. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, January 2017. 1–7. . https://doi.org/10.1109/ICACCS.2017.8014605

[2] Alaa AlSlaity, Banuchitra Suruliraj, Oladapo Oyebode, Jonathon Fowles, darren steeves, and Rita Orji. 2022. Mobile Applications for Health and Wellness: A Systematic Review. *Proc. ACM Hum.-Comput. Interact.* 6, EICS (June 2022), 171:1-171:29. https://doi.org/10.1145/3534525

[3] Andrea M. Barbarin, Laura R. Saslow, Mark S. Ackerman, and Tiffany C. Veinot. 2018. Toward Health Information Technology that Supports Overweight/Obese Women in Addressing Emotion- and Stress-Related Eating. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (*CHI '18*), April 21, 2018, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–14. . https://doi.org/10.1145/3173574.3173895

[4] Amit Baumel, Frederick Muench, Stav Edan, and John M. Kane. 2019. Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis. *Journal of Medical Internet Research* 21, 9 (September 2019), e14567. https://doi.org/10.2196/14567

[5] Claire L. Bentley, Lauren Powell, Stephen Potter, Jack Parker, Gail A. Mountain, Yvonne Kiera Bartlett, Jochen Farwer, Cath O'Connor, Jennifer Burns, Rachel L. Cresswell, Heather D. Dunn, and Mark S. Hawley. 2020. The Use of a Smartphone App and an Activity Tracker to Promote Physical Activity in the Management of Chronic Obstructive Pulmonary Disease: Randomized Controlled Feasibility Study. *JMIR mHealth and uHealth* 8, 6 (June 2020), e16203. https://doi.org/10.2196/16203

[6] Paul Bertens, Anna Guitart, and África Periáñez. 2017. Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, August 2017. 33–36. . https://doi.org/10.1109/CIG.2017.8080412

[7] Jonathan Bricker, Zhen Miao, Kristin Mull, Margarita Santiago-Torres, and David M. Vock. 2023. Can a Single Variable Predict Early Dropout From Digital Health Interventions? Comparison of Predictive Models From Two Large Randomized Trials. *J Med Internet Res* 25, (January 2023), e43629. https://doi.org/10.2196/43629

[8] David R. de Buisonjé, Fiona Brosig, Linda D. Breeman, Erika Litvin Bloom, Thomas Reijnders, Veronica R. Janssen, Roderik A. Kraaijenhagen, Harold M.C. Kemps, and Andrea W.M. Evers. 2023. Put your money where your feet are: The real-world effects of StepBet gamified deposit contracts for physical activity. *Internet Interv* 31, (February 2023), 100610. https://doi.org/10.1016/j.invent.2023.100610

[9] Eleanor R. Burgess, Elizabeth Kaziunas, and Maia Jacobs. 2022. Care Frictions: A Critical Reframing of Patient Noncompliance in Health Technology Design. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (November 2022), 281:1-281:31. https://doi.org/10.1145/3555172

[10] Grace H.Y. Chin and Kenny K.N. Chow. 2023. Technology-Enabled Interventions for Sustaining Behaviour Change in Adolescents: A Scoping Review for Research Gaps. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (September 2023), 1–30. https://doi.org/10.1145/3610211

[11] James Clawson, Jessica A. Pater, Andrew D. Miller, Elizabeth D. Mynatt, and Lena Mamykina. 2015. No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp '15*), September 07, 2015, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 647–658. . https://doi.org/10.1145/2750858.2807554

[12] Soumi De and P. Prabu. 2022. Predicting customer churn: A systematic literature review. *Journal of Discrete Mathematical Sciences and Cryptography* 25, 7 (October 2022), 1965–1985. https://doi.org/10.1080/09720529.2022.2133238

[13] Isuru Dharmasena, Mike Domaratzki, and Saman Muthukumarana. 2022. Comparison of Resampling Methods on Mobile Apps User Behavior. In *Internet of Things and Connected Technologies* (*Lecture Notes in Networks and Systems*), 2022, Cham. Springer International Publishing, Cham, 253–271. . https://doi.org/10.1007/978-3-030-94507-7_24

[14] Elizabeth V. Eikey and Madhu C. Reddy. 2017. "It's Definitely Been a Journey": A Qualitative Study on How Women with Eating Disorders Use Weight Loss Apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (*CHI '17*), May 02, 2017, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 642–654. . https://doi.org/10.1145/3025453.3025591

[15] Kamya Eria and Booma Poolan Marikannan. 2018. Systematic Review of Customer Churn Prediction in the Telecom Sector. 2, 1 (2018).

[16] Clemens Ernsting, Stephan U. Dombrowski, Monika Oedekoven, Julie L. O Sullivan, Melanie Kanzler, Adelheid Kuhlmey, and Paul Gellert. 2017. Using Smartphones and Health Apps to Change and Manage Health Behaviors: A Population-Based Survey. *J Med Internet Res* 19, 4 (April 2017), e101. https://doi.org/10.2196/jmir.6838

[17] Gunther Eysenbach. 2005. The law of attrition. *J Med Internet Res* 7, 1 (March 2005), e11. https://doi.org/10.2196/jmir.7.1.e11

[18] B.J. Fogg. 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[19] Jill Freyne, Ian Saunders, Emily Brindal, Shlomo Berkovsky, and Gregory Smith. 2012. Factors associated with persistent participation in an online diet intervention. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '12*), May 05, 2012, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 2375–2380. . https://doi.org/10.1145/2212776.2223805

[20] Aakash Ganju, Srini Satyan, Vatsal Tanna, and Sonia Rebecca Menezes. 2020. AI for Improving Children's Health: A Community Case Study. *Front Artif Intell* 3, (2020), 544972. https://doi.org/10.3389/frai.2020.544972

[21] David L. García, Àngela Nebot, and Alfredo Vellido. 2017. Intelligent data analysis approaches to churn as a business problem: a survey. *Knowl Inf Syst* 51, 3 (June 2017), 719–774. https://doi.org/10.1007/s10115-016-0995-z

[22] Louis Geiler, Séverine Affeldt, and Mohamed Nadif. 2022. An effective strategy for churn prediction and customer profiling. *Data & Knowledge Engineering* 142, (November 2022), 102100. https://doi.org/10.1016/j.datak.2022.102100

[23] Andrea Gentili, Giovanna Failla, Andriy Melnyk, Valeria Puleo, Gian Luca Di Tanna, Walter Ricciardi, and Fidelia Cascini. 2022. The cost-effectiveness of digital health interventions: A systematic review of the literature. *Frontiers in Public Health* 10, (2022). Retrieved September 13, 2023 from https://www.frontiersin.org/articles/10.3389/fpubh.2022.787135

[24] Drishti P. Ghelani, Lisa J. Moran, Cameron Johnson, Aya Mousa, and Negar Naderpoor. 2020. Mobile Apps for Weight Management: A Review of the Latest Evidence to Inform Practice. *Frontiers in Endocrinology* 11, (2020). Retrieved July 13, 2023 from https://www.frontiersin.org/articles/10.3389/fendo.2020.00412

[25] Priya Gopal and Nazri Bin MohdNawi. 2021. A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, December 2021. 1–8. . https://doi.org/10.1109/CSDE53843.2021.9718460

[26] Lisa M. Grega, Nadir Weibel, Shadia J. Assi, Natalie M. Golaszewski, Eric B. Hekler, and Job G. Godino. 2019. SMART 2.0: A Multimodal Weight Loss Intervention for Young Adults. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI EA '19*), May 02, 2019, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–6. . https://doi.org/10.1145/3290607.3312940

[27] Elina Helander, Kirsikka Kaipainen, Ilkka Korhonen, and Brian Wansink. 2014. Factors related to sustained use of a free mobile app for dietary self-monitoring with photography and peer feedback: retrospective cohort study. *J Med Internet Res* 16, 4 (April 2014), e109. https://doi.org/10.2196/jmir.3084

[28] Hugo Hesser. 2020. Estimating causal effects of internet interventions in the context of nonadherence. *Internet Interventions* 21, (September 2020), 100346. https://doi.org/10.1016/j.invent.2020.100346

[29] Robert Jakob, Samira Harperink, Aaron Maria Rudolf, Elgar Fleisch, Severin Haug, Jacqueline Louise Mair, Alicia Salamanca-Sanabria, and Tobias Kowatsch. 2022. Factors Influencing Adherence to mHealth Apps for Prevention or Management of Noncommunicable Diseases: Systematic Review. *Journal of Medical Internet Research* 24, 5 (May 2022), e35371. https://doi.org/10.2196/35371

[30] Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (*CHI '16*), May 07, 2016, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 2001–2013. . https://doi.org/10.1145/2858036.2858515

[31] Muhammad B. A. Joolfoo, Rameshwar A. Jugumauth, and Khalid M. B. A. Joolfoo. 2020. A Systematic Review of Algorithms applied for Telecom Churn Prediction. In *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, November 2020. 136–140. . https://doi.org/10.1109/ELECOM49001.2020.9296999

[32] Maurits Kaptein, Steven Duplinsky, and Panos Markopoulos. 2011. Means based adaptive persuasive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 07, 2011, Vancouver BC Canada. ACM, Vancouver BC Canada, 335–344. . https://doi.org/10.1145/1978942.1978990

[33] David Kerr, Fraya King, and David C. Klonoff. 2019. Digital Health Interventions for Diabetes: Everything to Gain and Nothing to Lose. *Diabetes Spectrum* 32, 3 (August 2019), 226–230. https://doi.org/10.2337/ds18-0085

[34] Adil Mehmood Khan and Seok-Won Lee. 2013. Need for a context-aware personalized health intervention system to ensure long-term behavior change to prevent obesity. In *2013 5th International Workshop on Software Engineering in Health Care (SEHC)*, May 2013, San Francisco, CA, USA. IEEE, San Francisco, CA, USA, 71–74. . https://doi.org/10.1109/SEHC.2013.6602481

[35] Ashik Khatri, Dvijesh Shastri, Panagiotis Tsiamyrtzis, Ilyas Uyanik, Ergun Akleman, and Ioannis Pavlidis. 2016. Effects of Simple Personalized Goals on the Usage of a Physical Activity App. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '16*), May 07, 2016, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 2249–2256. . https://doi.org/10.1145/2851581.2892366

[36] Seungwook Kim, Daeyoung Choi, Eunjung Lee, and Wonjong Rhee. 2017. Churn prediction of mobile and online casual games using play log data. *PLOS ONE* 12, 7 (July 2017), e0180735. https://doi.org/10.1371/journal.pone.0180735

[37] Rachel Kornfield, David C. Mohr, Rachel Ranney, Emily G. Lattie, Jonah Meyerhoff, Joseph J. Williams, and Madhu Reddy. 2022. Involving Crowdworkers with Lived Experience in Content-Development for Push-Based Digital Mental Health Tools: Lessons Learned from Crowdsourcing Mental Health Messages. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 99:1-99:30. https://doi.org/10.1145/3512946

[38] Sakris K. E. Kupila, Anu Joki, Laura-U. Suojanen, and Kirsi H. Pietiläinen. 2023. The Effectiveness of eHealth Interventions for Weight Loss and Weight Loss Maintenance in Adults with Overweight or Obesity: A Systematic Review of Systematic Reviews. *Curr Obes Rep* (June 2023). https://doi.org/10.1007/s13679-023-00515-2

[39] Hongwook Kwon, Ho Heon Kim, Jaeil An, Jae-Ho Lee, and Yu Rang Park. 2021. Lifelog Data-Based Prediction Model of Digital Health Care App Customer Churn: Retrospective Observational Study. *J Med Internet Res* 23, 1 (January 2021), e22184. https://doi.org/10.2196/22184

[40] Joyca Lacroix, Privender Saini, and Annelies Goris. 2009. Understanding user cognitions to guide the tailoring of persuasive technology-based physical activity interventions. In *Proceedings of the 4th International Conference on Persuasive Technology* (*Persuasive '09*), April 26, 2009, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–8. . https://doi.org/10.1145/1541948.1541961

[41] Tricia Leahey and Jamie Rosen. 2014. DietBet: A Web-Based Program that Uses Social Gaming and Financial Incentives to Promote Weight Loss. *JMIR Serious Games* 2, 1 (February 2014), e2. https://doi.org/10.2196/games.2987

[42] EunJo Lee, Yoonjae Jang, DuMim Yoon, JiHoon Jeon, Seong-il Yang, Sang-Kwang Lee, Dae-Wook Kim, Pei Pei Chen, Anna Guitart, Paul Bertens, África Periáñez, Fabian Hadiji, Marc Müller, Youngjun Joo, Jiyeon Lee, Inchon Hwang, and Kyung-Joong Kim. 2019. Game Data Mining Competition on Churn Prediction and Survival Analysis using Commercial Game Log Data. *IEEE Trans. Games* 11, 3 (September 2019), 215–226. https://doi.org/10.1109/TG.2018.2888863

[43] Robert Lewis, Craig Ferguson, Chelsey Wilks, Noah Jones, and Rosalind W. Picard. 2022. Can a Recommender System Support Treatment Personalisation in Digital Mental Health Therapy? A Quantitative Feasibility Assessment Using Data from a Behavioural Activation Therapy App. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (*CHI EA '22*), April 28, 2022, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–8. . https://doi.org/10.1145/3491101.3519840

[44] Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. 2020. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1 (March 2020), 18:1-18:22. https://doi.org/10.1145/3381007

[45] Yevgeniy Medynskiy, Svetlana Yarosh, and Elizabeth Mynatt. 2011. Five strategies for supporting healthy behavior change. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '11*), May 07, 2011, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1333–1338. . https://doi.org/10.1145/1979742.1979770

[46] Jonah Meyerhoff, Rachel Kornfield, David C. Mohr, and Madhu Reddy. 2022. Meeting Young Adults' Social Support Needs across the Health Behavior Change Journey: Implications for Digital Mental Health Tools. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (November 2022), 312:1-312:33. https://doi.org/10.1145/3555203

[47] Gideon Meyerowitz-Katz, Sumathy Ravi, Leonard Arnolda, Xiaoqi Feng, Glen Maberly, and Thomas Astell-Burt. 2020. Rates of Attrition and Dropout in App-Based Interventions for Chronic Disease: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research* 22, 9 (September 2020), e20283. https://doi.org/10.2196/20283

[48] Elliot G. Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N. Tobin, Patricia Davidson, Arlene M. Smaldone, and Lena Mamykina. 2021. Automated vs. Human Health Coaching: Exploring Participant and Practitioner Experiences. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 99:1-99:37. https://doi.org/10.1145/3449173

[49] Vasiliki Mylonopoulou. 2018. Design for health behavior change supportive technology: healthcare professionals' perspective. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (*NordiCHI '18*), September 29, 2018, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 82–92. . https://doi.org/10.1145/3240167.3240196

[50] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2022. Systematic Review on Churn Prediction Systems in Telecommunications. In *Proceedings of Third International Conference on Communication, Computing and Electronics Systems* (*Lecture Notes in Electrical Engineering*), 2022, Singapore. Springer, Singapore, 983–995. . https://doi.org/10.1007/978-981-16-8862-1_64

[51] Nadeem Naz, Umar Shoaib, and M. Sarfraz. 2018. A REVIEW ON CUSTOMER CHURN PREDICTION DATA MINING MODELING TECHNIQUES. *Indian Journal of Science and Technology* 11, (July 2018), 1–7. https://doi.org/10.17485/ijst/2018/v11i27/121478

[52] Seth M. Noar, Christina N. Benac, and Melissa S. Harris. 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychol Bull* 133, 4 (July 2007), 673–693. https://doi.org/10.1037/0033-2909.133.4.673

[53] Heather L. O'Brien, Ido Roll, Andrea Kampen, and Nilou Davoudi. 2022. Rethinking (Dis)engagement in human-computer interaction. *Computers in Human Behavior* 128, (March 2022), 107109. https://doi.org/10.1016/j.chb.2021.107109

[54] Babaniyi Yusuf Olaniyi, Ana Fernández del Río, África Periáñez, and Lauren Bellhouse. 2022. User Engagement in Mobile Health Applications. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (*KDD '22*), 2022, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 4704–4712. . https://doi.org/10.1145/3534678.3542681

[55] Miranda Olff. 2015. Mobile mental health: a challenging research agenda. *Eur J Psychotraumatol* 6, (January 2015), 27882. https://doi.org/10.3402/ejpt.v6.27882

[56] Rita Orji, Regan L. Mandryk, Julita Vassileva, and Kathrin M. Gerling. 2013. Tailoring persuasive health games to gamer type. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '13*), April 27, 2013, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 2467–2476. . https://doi.org/10.1145/2470654.2481341

[57] Rita Orji, Lennart E. Nacke, and Chrysanne Di Marco. 2017. Towards Personality-driven Persuasive Health Games and Gamified Systems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (*CHI '17*), May 02, 2017, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1015–1027. . https://doi.org/10.1145/3025453.3025577

[58] Rita Orji, Gustavo F. Tondello, and Lennart E. Nacke. 2018. Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (*CHI '18*), April 21, 2018, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–14. . https://doi.org/10.1145/3173574.3174009

[59] Oladapo Oyebode, Chinenye Ndulue, Dinesh Mulchandani, Ashfaq A. Zamil Adib, Mona Alhasani, and Rita Orji. 2021. Tailoring Persuasive and Behaviour Change Systems Based on Stages of Change and Motivation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (*CHI '21*), May 07, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–19. . https://doi.org/10.1145/3411764.3445619

[60] Kiemute Oyibo and Julita Vassileva. 2020. Persuasive Features that Drive the Adoption of a Fitness Application and the Moderating Effect of Age and Gender. *MTI* 4, 2 (May 2020), 17. https://doi.org/10.3390/mti4020017

[61] Joonyoung Park and Uichin Lee. 2023. Understanding Disengagement in Just-in-Time Mobile Health Interventions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2 (June 2023), 72:1-72:27. https://doi.org/10.1145/3596240

[62] Daniel Hansen Pedersen, Marjan Mansourvar, Camilla Sortsø, and Thomas Schmidt. 2019. Predicting Dropouts From an Electronic Health Platform for Lifestyle Interventions: Analysis of Methods and Predictors. *J Med Internet Res* 21, 9 (September 2019), e13617. https://doi.org/10.2196/13617

[63] Tristan J. Philippe, Naureen Sikder, Anna Jackson, Maya E. Koblanski, Eric Liow, Andreas Pilarinos, and Krisztina Vasarhelyi. 2022. Digital Health Interventions for Delivery of Mental Health Care: Systematic and Comprehensive Meta-Review. *JMIR Mental Health* 9, 5 (May 2022), e35159. https://doi.org/10.2196/35159

[64] Charlie Pinder, Jo Vermeulen, Benjamin R. Cowan, and Russell Beale. 2018. Digital Behaviour Change Interventions to Break and Form Habits. *ACM Trans. Comput.-Hum. Interact.* 25, 3 (June 2018), 15:1-15:66. https://doi.org/10.1145/3196830

[65] J. O. Prochaska and W. F. Velicer. 1997. The transtheoretical model of health behavior change. *Am J Health Promot* 12, 1 (1997), 38–48. https://doi.org/10.4278/0890-1171-12.1.38

[66] Stephen Purpura, Victoria Schwanda, Kaiton Williams, William Stubler, and Phoebe Sengers. 2011. Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '11*), May 07, 2011, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 423–432. . https://doi.org/10.1145/1978942.1979003

[67] Amon Rapp, Maurizio Tirassa, and Lia Tirabeni. 2019. Rethinking Technologies for Behavior Change: A View from the Inside of Human Change. *ACM Trans. Comput.-Hum. Interact.* 26, 4 (June 2019), 22:1-22:30. https://doi.org/10.1145/3318142

[68] Kathleen Ryan, Samantha Dockray, and Conor Linehan. 2022. Understanding How eHealth Coaches Tailor Support For Weight Loss: Towards the Design of Person-Centered Coaching Systems. In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022, New Orleans LA USA. ACM, New Orleans LA USA, 1–16. . https://doi.org/10.1145/3491102.3501864

[69] Miro Schleicher, Rüdiger Pryss, Winfried Schlee, and Myra Spiliopoulou. 2022. When Can I Expect the mHealth User to Return? Prediction Meets Time Series with Gaps. In *Artificial Intelligence in Medicine* (*Lecture Notes in Computer Science*), 2022, Cham. Springer International Publishing, Cham, 310–320. . https://doi.org/10.1007/978-3-031-09342-5_30

[70] Victoria Schwanda, Steven Ibara, Lindsay Reynolds, and Dan Cosley. 2011. Side effects and "gateway" tools: advocating a broader look at evaluating persuasive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 07, 2011, Vancouver BC Canada. ACM, Vancouver BC Canada, 345–348. . https://doi.org/10.1145/1978942.1978991

[71] James Shaw, Payal Agarwal, Laura Desveaux, Daniel Cornejo Palma, Vess Stamenova, Trevor Jamieson, Rebecca Yang, R. Sacha Bhatia, and Onil Bhattacharyya. 2018. Beyond "implementation": digital health innovation and service design. *npj Digital Med* 1, 1 (September 2018), 1–5. https://doi.org/10.1038/s41746-018-0059-8

[72] Susan M. Shortreed, Eric Laber, Daniel J. Lizotte, T. Scott Stroup, Joelle Pineau, and Susan A. Murphy. 2011. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Mach Learn* 84, 1 (July 2011), 109–136. https://doi.org/10.1007/s10994-010-5229-0

[73] Pedro Sobreiro, Domingos Dos Santos Martinho, Jose G. Alonso, and Javier Berrocal. 2022. A SLR on Customer Dropout Prediction. *IEEE Access* 10, (2022), 14529–14547. https://doi.org/10.1109/ACCESS.2022.3146397

[74] Anika Steinert, Cornelia Eicher, Marten Haesner, and Elisabeth Steinhagen-Thiessen. 2020. Effects of a long-term smartphone-based self-monitoring intervention in patients with lipid metabolism disorders. *Assistive Technology* 32, 2 (March 2020), 109–116. https://doi.org/10.1080/10400435.2018.1493710

[75] Callie Thomson, Jane Nash, and Anthony Maeder. 2016. Persuasive Design for Behaviour Change Apps: Issues for Designers. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists* (*SAICSIT '16*), September 26, 2016, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–10. . https://doi.org/10.1145/2987491.2987535

[76] Helma Torkamaan and Jürgen Ziegler. 2021. Integrating Behavior Change and Persuasive Design Theories into an Example Mobile Health Recommender System. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (*UbiComp/ISWC '21 Adjunct*), September 24, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 218–225. . https://doi.org/10.1145/3460418.3479330

[77] Anna L. Trella, Kelly W. Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A. Murphy. 2022. Designing Reinforcement Learning Algorithms for Digital Interventions: Pre-Implementation Guidelines. *Algorithms* 15, 8 (August 2022), 255. https://doi.org/10.3390/a15080255

[78] Ha Trinh, Ameneh Shamekhi, Everlyne Kimani, and Timothy W. Bickmore. 2018. Predicting User Engagement in Longitudinal Interventions with Virtual Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (*IVA '18*), 2018, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 9–16. . https://doi.org/10.1145/3267851.3267909

[79] Chih-Fong Tsai and Yu-Hsin Lu. Data Mining Techniques in Customer Churn Prediction.

[80] Kelsey Ufholz and James Werner. 2023. The Efficacy of Mobile Applications for Weight Loss. *Curr Cardiovasc Risk Rep* 17, 4 (April 2023), 83–90. https://doi.org/10.1007/s12170-023-00717-2

[81] V Umayaparvathi and K Iyakutti. A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. 03, 04 .

[82] Youfa Wang, Hong Xue, Yaqi Huang, Lili Huang, and Dongsong Zhang. 2017. A Systematic Review of Application and Effectiveness of mHealth Interventions for Obesity and Diabetes Treatment and Self-Management. *Adv Nutr* 8, 3 (May 2017), 449–462. https://doi.org/10.3945/an.116.014100

[83] R. Jay Widmer, Nerissa M. Collins, C. Scott Collins, Colin P. West, Lilach O. Lerman, and Amir Lerman. 2015. Digital Health Interventions for the Prevention of Cardiovascular Disease: A Systematic Review and Meta-analysis. *Mayo Clinic Proceedings* 90, 4 (April 2015), 469–480. https://doi.org/10.1016/j.mayocp.2014.12.026

[84] Markus Zanker, Laurens Rook, and Dietmar Jannach. 2019. Measuring the impact of online personalisation: Past, present and future. *International Journal of Human-Computer Studies* 131, (November 2019), 160–168. https://doi.org/10.1016/j.ijhcs.2019.06.006

[85] Jichen Zhu, Diane H. Dallal, Robert C. Gray, Jennifer Villareale, Santiago Ontañón, Evan M. Forman, and Danielle Arigo. 2021. Personalization Paradox in Behavior Change Apps: Lessons from a Social Comparison-Based Personalized App for Physical Activity. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 116:1-116:21. https://doi.org/10.1145/3449190

[86] Non communicable diseases. Retrieved July 13, 2023 from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[87] Organic and paid app installs retention rate 2020. *Statista*. Retrieved September 4, 2023 from https://www.statista.com/statistics/1263373/app-retention-organic-paid-installs/