USING ARTIFICIAL INTELLIGENCE TO PERSONALIZE MOBILE APPLICATIONS

Brazhnykov Andrii Mykolaiovych [1]

Received: 2025-09-10
Accepted: 2025-10-20
DOI: https://doi.org/10.5281/zenodo.18705392

Abstract. The article is an investigation into the current state of the art in applying AI techniques for personalizing mobile applications in the course of the evolution of mobile computing platforms and increasing requirements on confidentiality and adaptability of digital services. The author highlights the need for a personal user experience based on individual behavioral profiles, context information, scenarios of local interactions, and device constraints. Existing techniques for mobile personalization, such as machine learning and deep learning approaches, recommendation engines, NLP, real, time analytics, and hybrid methods of adapting content are reviewed.

Much attention is given to the issues of performance and resource limitations of mobile systems, and specifics of local model optimization. The impact of the hardware structure (processors) on the inference speed, energy efficiency, stability of personalization features in on, device mode is studied. It is demonstrated that it is precisely the presence of specialized neural processors that allows running a complex AI model without using cloud technology and at low processing latency, guaranteeing data security.

The paper talks about privacy problems and study the privacy, preserving personalization framework such as federated learning and differential privacy. It has been proven that the aggregation of local learning, local noisy mixing mechanism and optimization of neural models can provide private personalized services. The characteristics of the new generation of mobile artificial intelligence ecosphere is considered and their use for integrating personalized functionality, fast local model execution and creating adaptive user interface is identified.

Keywords: personalization, mobile applications, machine learning, artificial intelligence, neural networks, mobile ecosystems.

[1] Master's Degree in Specialized Computer Systems,
Kharkiv National University of Radioelectronics, Ukraine
brazhnicov@gmail.com
https://orcid.org/0009-0000-4066-757X

Introduction

The rapid adoption of the new application platforms led to radical changes in both amount of consumed content/services, and in ways of its usage, thereby influencing individual industries, and user behavior. Ultraflexible mobile platforms grew into big ecosystems providing users access to a wide variety of e, commerce services, entertainments, education, healthcare, and many other service categories for an equally diverse number of needs [2]. Ironically, open, ended nature of existing OS/devices, although all inclusive, still appears to short term, since users get easily bored with offered service and not necessarily interested to individual needs, desires and motivations [1].

With this problem in sight, personalization receives more interest from both researchers and practitioners and grows into an important attribute in enhance perceived quality of UX design and the retention of loyal users. Mobile application personalization (MAP) can be interpreted as tailoring application UI, content and operational aspects to user's need and preference, resulting in improved user satisfaction and retention [2]. AI based personalization (AIBP) is fundamentally different from classical and rule, based approaches in that intelligent methods utilized ML and DL algorithms to predict and detect user interest and behavior and adjust application functionalities and services correspondingly in proactive manner [1, 3]. With explicit user preferences such as ratings, comments, review and favor list as well as implicit behavioral factors including clicks, view, interactions, time and transition filters. This intelligent solution combines together various kind of user preferences providing the personalization context and rich user experiences in real time [4].

The additional capabilities of modern mobile platform, such as on, device computing, sensors and geo, tagged data, make the personalized recommendations extremely relevant and context, aware. Smart applications can combine time, location and interaction dynamics data with other signals from fine, grained context, producing smart and readily available personalized output [1, 5]. Some authors highlight the beneficial influences that such system provides, enhancing customer and amateur user`s engagement and trust and increasing conversion rates for e, commerce and social networks. AI, powered business, consumer and Internet services (such as smart predictions or context awareness systems and enriched system with social communities) improve users' experience and engagement [6].

Despite all these benefits, the automatisation of mobile app personalization system with AI raises a practical and ethical dilemma concerning the protection of user data, personal privacy and transparence thus leading to practical, ethical and operational problems. Leaks in information, personal data security or violation may happen while the construction of user interest graph and cognitive map. Excessive informative personalization system may also affect user engagement and trust and will not favor the relation between the user and the service. In spite of the desire to have a personalization without being too intrusive, a skeptical audience may not give favorable reviews. Solutions such as federated learning or differential privacy may limit these shortcomings, however finding a balance between the tailored user experience and user privacy is a current problem.

Statement of the problem. The scientific and technological problems caused by the use of AI for personalized mobile applications, although AI has already gained tremendous success in development of AI. We focus on:

Resource limitations. The limited hardware resource of mobile devices (low CPU, memory and GPU; the huge energy consumption of the battery, etc.) can be a deterrent in deploying machine learning models of higher orders. The solution are the various available AI and hardware, focused platforms and AI frameworks (TensorFlow Lite, Core ML, NPU based, oriented frameworks) which aim to minimize memory and computing load. The difficulty lies in devising models that meet user expectations with less computational resources.

User data privacy and security. With regard to the collection and processing of mobile application personalization data, spanning behavioral, contextual, and sensory information, sensitive personal data will be inevitably exposed. Measures must be taken to avoid data theft, leaks, and misuse of data, with an optimized AI solution as content personalization. In particular, minimal training data sets should be used, On, device processing will have to be prioritized, communication between parties should be minimized. International standards regarding information safety should be complied.

Variability of mobile platform ecosystems or operating systems (OS). The native or on, cloud mode of OS operation in mobile devices limit the use of a general approach to personalization. Since Android, based and iOS, based mobile devices have varied OS infrastructure, development toolkits and hardware support of AI, ecosystem operators present their own implementations (Galaxy AI, Xiaomi Hyper AI, Moto AI, Apple Intelligence etc.) for models, which result in diversity in the service implementation for personalization. Creating a generalized but still feasible solution for multiple OS seems to be problematic.

User behavior. Fast changes and oscillation in users' attention levels make it challenging to build stable personalization. Varying user preference across various contexts, surroundings, seasons, days, time of seconds, and degrees of interaction demands models that can not only predict future user trajectories but also enrich the model's learning levels in retraining mode. Model parameters (weights) update must be automatic and continuous.

Analysis of recent studies and publications. The area where the application of AI in the applications (or apps) of mobiles has become most pertinent in contemporary research is [1], since with growing number of functions of modern systems mobile intelligence is needed for adjustment, optimization and personalization. The theoretical basis and methodologies for AI application to mobile systems explored in major modern scientific works were formed around the principle of construction of intelligent applications for mobile devices where context analysis of user activity is crucial. In particular, [7, 8] serve as a foundation for a theoretical framework presenting the benefits of using rule based adaptive management systems and explaining the increased precision when utilizing contextual information (location, user activity, past user interaction), while discussing challenges of scaling expert systems in multi component mobile applications.

Significant work has been conducted in the area of personalization of mobile experience; the literature [9, 10] which is referred supports this position of the interesting phenomenon of moving from responsive interface to what is now being referred as 'predict and adapt', where the application would automatically alter its configuration in the user, defined requirements. In addition to behavioral profiling and hybrid recommender system, which achieves greater than 80% accuracy of what the configuration should be for service provision computational cost issue and data privacy risk are also mentioned by the authors.

Secondly, the equally important dimension of AI, where personalization of mobile applications is through intelligent management of the applications and resources. The topic of performance of mobile applications in relation to AI is tackled in [11] and presented is load prediction based optimization, sliding window resource allocation and energy efficient process management. The researchers show that on the area of mobile applications, intelligent approaches can lead to lower delays, reduced memory consumption and higher application reliability. The limitation posed by finite resources on mobile devices, is the overall common challenge, which can either be resolved through lightweight machine learning models or dynamic system architectures. Almost at the same time in [12] an author gives a systematic review of using AI methods in real mobile systems, recognizing similar methods and online learning, coefficient estimation of regression models and context aware models as the most efficient. They point out that hybrid solutions present the best trade, off between performance, accuracy and efficiency, yet mobile device resource limitation is the core issue preventing the usage of advanced deep learning techniques.

The discussed models for enabling access to mobile app with AI, for instance [13, 14], should take into consideration also the access to AI and common mobile UI. The authors mention that intelligent services with automatic element scaling, voice navigations, and adaptable layout may lead to universal mobile platforms for users with all kinds of disabilities. However, the standardization and compatibility issues with different UI frameworks have not been fully solved.

Issue on Usability Evaluation with ML is addressed in [15]. Here, researchers are tackling issues for predicting usability or evaluate the interface with the help of AI. They propose to use Large UI Models for predicting UI or for automating UI analysis on applications. In this context, Big UI models could be used to perform predictions and evaluate interface design based on example UI interactions. One obstacle found by researchers here is the scarcity of robust UI pattern collections and poor models in modeling real life situations.

Some data about the applied and industrial investigations with regard to the AI research and development trends is shown in the sources [16, 17], review materials with the applied nature of implementation of the AI into mobile development are published there. In the sources authors especially focus on application of the NLP, chatbots, computer vision methods, personalization systems, predictive analytics, etc. Though the level of the scientific support is lower these publications enable understanding of current trends in industry and the global directions of the functional implementation.

Another study [18] also presents the application of AI in infrastructure optimization and technical processes, in particular by considering the use of machine learning for the network devices configuration management. Even if it is not directly related to mobile application it proves that AI can lead to great potential in the technical automation that can be exploited in mobile application in order to offer sophisticated network services.

From the review of prior up to date works the problem of using AI for mobile applications and app customization is that there is not one single architecture that would embrace the dimensions such as personalization, usability analysis, performance tuning and inclusiveness all at the same time within a single framework. Most of the research is focused on only a part of these features so for a system to take advantage of all of these features, a cross, disciplinary approach of designing context, aware, light, weight and scalable models has to be employed. However, this justifies an availability of the scientific works for an all, round AI architectures for mobile apps which performs well under limited resources.

Aim and tasks of the study. Aim of the study is to provide an in, depth investigation of methods of mobile applications personalization based on artificial intelligence, and assess possibilities of modern mobile computing platforms to provide an efficient, privacy and context, sensitive adaptation of user experience under resource constraints. To achieve this aim, the tasks have been specified:

Analyze present day technological ways toward on, line customization of service and software for mobile telecoms, such as the machine learning approach, deep learning algorithms, NLP and real, time prognostics and flexible situational models;

Explore the critical issues of AI implementation in the mobile environment including the limitation of resources within mobile devices, data privacy, operating system's heterogeneity, and the fluctuating user behavior.

Evaluate secure personalized models (federated learning, differential privacy) which can protect the personalization of mobile apps

To classify and analyze the properties of computing modules of modern mobile platforms and their effect on the speed and quality of personalization of a mobile device.

Examine ecosystems, looking at how well they map onto smartphone hardware features; and

Results

Models of machine and deep learning, approaches of natural language processing, as well as real time analytical system mechanics, contextual approaches have been systematized. Next part connects to the analysis of private personalization systems (including federated learning and differential privacy) then focuses on the research of hardware capacities of present, day mobile devices (including neural processing units research, computing accelerators and AI systems). Such order provides smooth development from methodological assumptions toward technical realization and hardware basis of personalized systems.

Mobile application AI personalization techniques: Artificial Intelligence is employed on the mobile applications, the AI technology used for the mobile application personalization makes it more user centric. The mobile applications leverage the AI techniques namely ML, DL, NLP, real time analytics, privacy aware frame work and etch for the purpose of personalization the content, application and features which increases the visits, engagement and retention of the user [7, 10].

ML-based approaches. ML is an important technology to use in mobile applications to personalize its system. It provides you to analyze the user data on a large-scale basis and extract hidden patterns within user`s behaviors that cannot be found by manual processing. Because ML helps to build systems that can adapt to and predict user preferences, personalize interfaces and content and respond dynamically to user`s behavior changes. One of the most traditional techniques of personalizing a system, Collaborative Filtering (CF), suggests that users with similar preferences will react similarly to items. From this principle we can deduce that a user`s preferences will follow the user history of actions performed or a relation between the items. Let's consider the main approaches of CF [10].

User-based collaborative filtering. An algorithm takes into consideration a community of users and identify the users that have similar behavior to the target user. Then the content that similar community liked is recommended to the target user. (shown in Fig. 1) for instance, in streaming music service (Spotify) the application recommendation the music which is loved by users with similar preference, in mobile education applications, courses that have similar learning interests are recommended to students. Pros of the approach are simple and efficient use of group knowledge while Cons include heavy dependence on the behavior data of other users and cold start for new user.

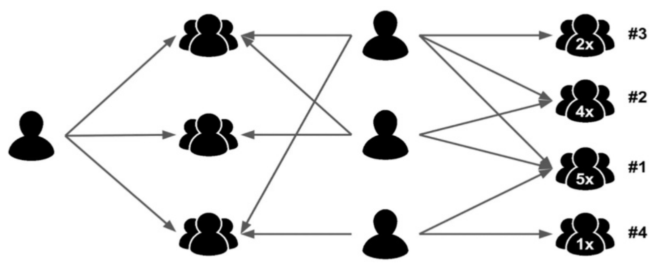

Figure 1 - User-based collaborative filtering algorithm

Source: created by the author based on [19]

Joint filtering based on elements. In this method, the relations between the contents are analyzed, where the system figures out which items will probably be used or rated together and will recommend similar items to the user. For example, in an online shop where "A" and "B" were commonly purchased together, if the user chose "A", it will recommend "B". For a media service, it is a recommendation of a movie or article based on viewing similar user's behaviors, movie recommendation on Netflix for instance as shown in Figure 2. Advantages include low dependency to the number of users and it is efficient for a large content dataset;

Disadvantages include that individual users' attributes are not always taken into consideration, and it can't always be adapt to user's changing interests.

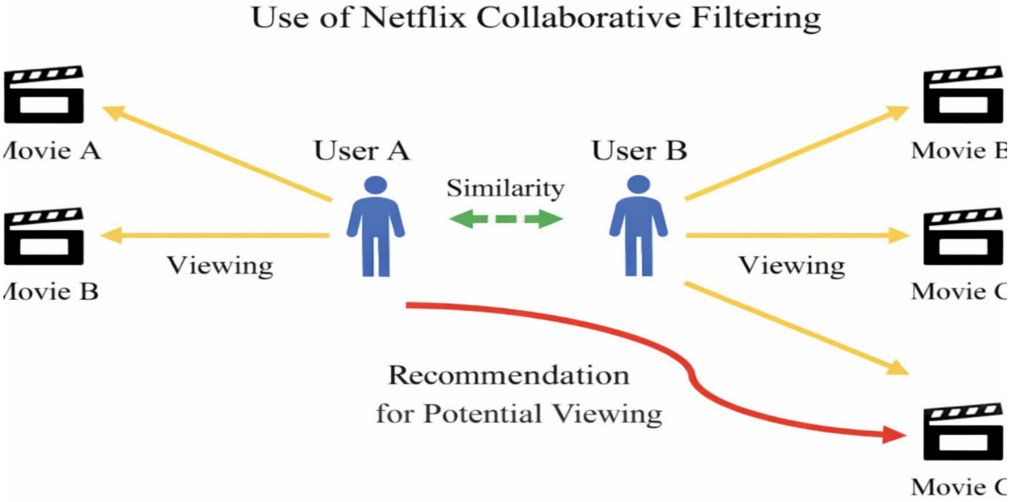

Figure 2 - Using AI for mobile app personalization with collaborative filtering, using Netflix as an example

Source: created by the author based on [19]

Hybrid models. Nowadays the personalization systems are developing using hybrid models that integrate the common filtering approaches and content-based methods. These models take into account user's preferences (based on his behaviors) and the features of content (using content filtering) that lead to: better recommendation accuracy, elimination of "cold" start problem, higher variety of content, integration of other contexts (time, place, device, etc.). An example: In a streaming service, we have user's preferences in some genres of music (using content filtering) and we consider characteristics of songs (mood, genre, etc.). Using these two kinds of information, system can suggest a playlist for the user.

Deep learning-based techniques. The application of deep learning enables personalization for mobile applications to a further extent because deep learning models can represent non-linear relationships of varying complexity among users and content. By applying multilayer neural networks, the system is not just able to analyze the history of user actions but can also take user context, time, content types, etc into account to capture user interests [8, 10].

Neural models for recommendation: Currently, neural networks of all kinds are quite commonly used for personalization: CNN, RNN, Transformer, see Fig. 3:

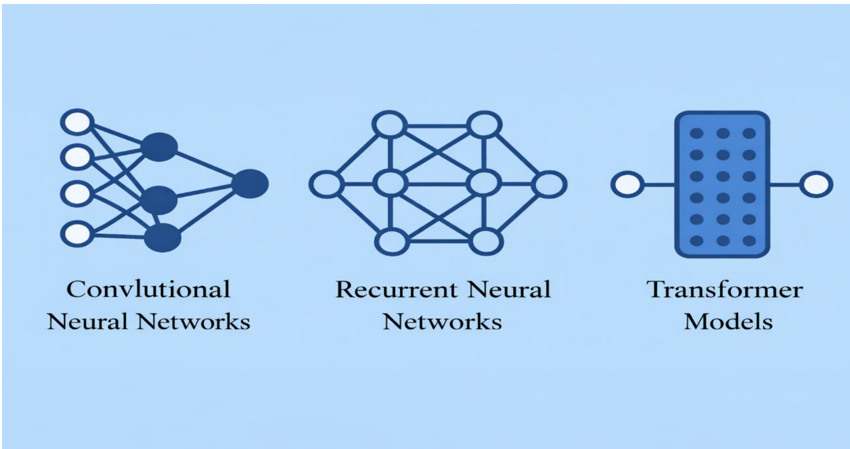

Figure 3 - Neural architectures of ZNM, RNM, and transformed models for mobile app personalization

Source: compiled by the author

Convolutional neural networks (CNN)s that work well with content that has a spatial relationship such as images, videos. For personalization, CNN can find the content attributes that appeal to users. For example, in a mobile shopping app CNNs analyze images of the product for making similar product recommendation or offering a replacement that fits with the user's style and preference [7];

Recurrent neural networks (RNNs) are effective for next item/rating prediction as they can take into account a user's sequential behavior over time, that is, changes in user preferences (e. g. In streaming applications for video or music, RNNscan predict the next video the user will watch based on the previous sessions) [8];

Transformed models are efficient in analyzing long sequence data, and modeling dependency between user actions. For example, when providing personalization services for educational platform or social network, they are very efficient. The models above analyze student's learning behavior datahistory, to predict relevant courses or resources to the student [3, 9].

The strengths of deep learning in terms of mobile app personalization is: extracting of the complex and hidden relations between data; capability to be applied to diverse kinds of content (text, image, video and sound); continuous adaptation of recommendation based on the user behaviors over time. The weaknesses for the ML techniques: a high demand for computing resource; demand for large quantity of data; explain and interpret of the model result.

Personalization based on natural language processing (NLP). Using NLP techniques personalization systems can process text-based information and understand the meaning and contents of the message, comment, review or user request (Fig. 4). This further aids in personalized adaptation of content and personalized anticipation of the user requirements. NLP applications include; education systems using student feedback written on text reviews and questions and deriving personalized learning plans; mobile news application delivering user personalized news feed using analysis of the meaning/ semantic content of news article read by user; user support service utilizing OPM to build an interactive chat agent which understands the context of a user request and then gives him a personalized recommendation. In the figure 4 you can see OPM's relationship in personalization and recommendation systems for personalization on a mobile application [1, 10, 20].

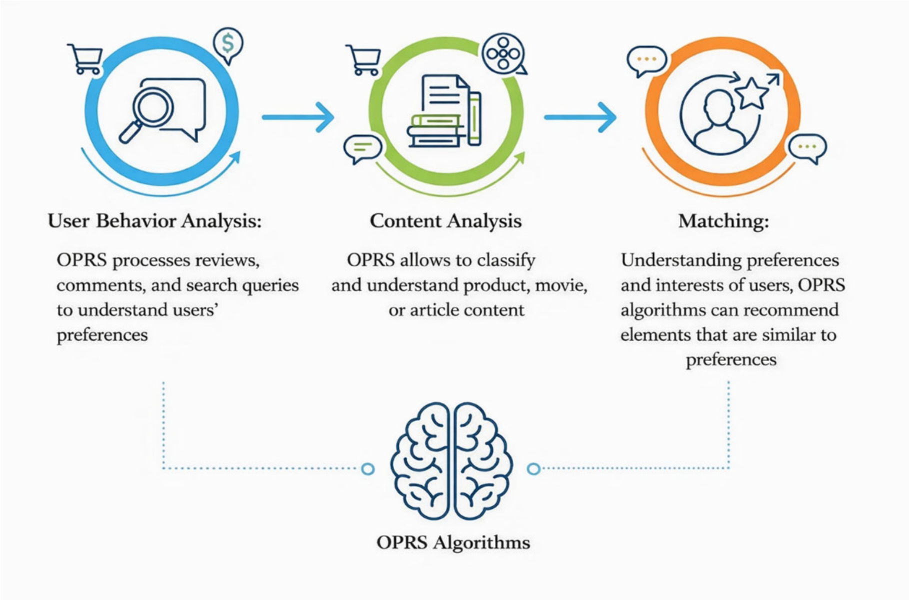

Figure 4 - Use of NLP in personalization and recommendation systems for mobile app personalization

Source: created by the author based on [20]

The benefits of NLP include the ability to more accurately personalize text content (understanding users' intentions and interests) and incorporate contextual and behavioral data into recommendations. The disadvantages of personalization using NLP are, in summary, complexity in dealing with multi-lingual text, large corpora required to train a model, and bias of the model towards user data.

Real time data analytics (RTDA) and Contextual Analytics (CA). One of the main part of mobile personalization is to provide users with dynamically tailored content and functions, which RTDA takes into consideration in real-time [10]. Traditionally, analytics takes historical data to develop predictions, recommendations after a while, while real-time system factors in events that have been taking place, users' current actions and environmental issues.

Real-time data analytics empowers the applications to react in real time in reaction to user actions and changes in the context, through the inference techniques (estimation of coefficients based on some measurements/data) with low latency, model compression, quantization, and computation on the device. Applications include: transportation applications (application analyzes route, speed, and traffic of the user, to generate and adapt personalized recommendations about most convenient route/delivery time) streaming services for mobile devices (during playback of video or audio application, a user may receive personalized playlist, or song suggestions based on session, user behavior) E-commerce services for mobile devices, on user's actions (views, clicks, add-to-cart actions) application can send user the personalized offers or alternative products. The benefits of this approach: real time response for new context, improvement of user satisfaction and user engagement, better adaptation to dynamic environment of interaction with application, while problems associated: demands on computation and energy, difficulties in integration with models, difficulties in integration with existing application infrastructure, and management of data flow to eliminate delays.

Contextual analytics enables the personalization to be influenced by external and internal aspects which can affect the personalized experience. This will be user's location, time, device that is being used, ambient conditions, sensor data etc. Contextual analytics example: Geolocation services (recommendations of a restaurant or shop based on the location and current time the user entered it), Fitness applications (personalized fitness or training plans based on the user's location, available equipment and activity habits, as well as weather data and other factors), E-learning (personalization of courses or exercises based on the learning habits of the user, the time of the day they interact the most with the system), etc. Benefits of contextual analytics: personalization is more obvious, more interpretable, and more relevant. The recommendation effectivity is increased by considering the real context, user conversion and loyalty also increase.

4. Personalization frameworks that preserve confidentiality.

Personalization is becoming one of the key requirements in modern mobile applications, in which there is a need to simultaneously provide accurate recommendations while respecting data privacy and security constraints. The increased regulatory requirements, like GDPR and CCPA, together with user concerns about their private data, caused to appear new methods to build personalized systems reducing the possibility of data leak and leaving user control over data [10].

There are two main approaches that are widely recognized as frameworks for private personalization, i. e. Federated Learning (FL) and Differentially Privacy (DP), while combining FL with DP has been emerging as the de facto approach for private personalization.

FL is a decentralized method for ML model training that allows personalization of a mobile application without sending private user data to the server. FL works by first: 1) The model is sent to the device from the server. 2) The device learns a local model by using data stored locally. 3) Only the weights (gradients) are sent to the server from the device and not the data. 4) The server combines all the updates to create a better global model. 5) The process is repeated. The benefits of FL are total preservation of user privacy and secrecy, model personalization of the individual device, lesser chance of massive data leakage and lesser burden on the cloud infrastructure.

Let's look at a few examples and trends in FL:

- mobile keyboards (Gboard) adapt autocorrect and suggestions;
- fitness apps customize individual training plans;
- medical mobile apps personalize recommendations without transferring medical data;
- partial model personalization, where some layers remain local;
- FL based on noisy updates for better privacy;
- combining FL with contextual signals to increase adaptability.

DP is a mathematically proved technique which assures that it is impossible to reveal that an individual user was participating in the data set, given that the outcome of the whole computation is revealed to a third party. Using noisy training data models are achieved high levels of privacy, without losing accuracy when forecasting. The concept of DP is that: 1) the statistical noise added to the data or model updates does not depend on which data point contributes to the model; 2) adding or removing data points should not substantially alter the output; 3) it should be impossible to determine if a specific user participated in the dataset based on aggregated outputs of statistics or analysis. The features of DP include high mathematically proved privacy level, it permits statistics and analysis without the danger of deanonymization, it works together with FL and other environments. DP is utilized in: the educational systems (where aggregated data about academic progress can be collected), mobile analytics tools (where data can be collected without personal identifiers), big technology organizations like Apple or Google (for collection of telemetry data in Android and iOS systems) [5, 10].

Computing resources for AI-based personalization and ecosystems for mobile app personalization in modern devices.

Mobile app personalization using AI requires a lot of computing resources as ML and DL algorithms deal with multi, dimensional data streams, look for behavioural patterns, and execute model inference in real time. The performance and speed of personalized features directly depend on hardware components available to the mobile device and how well optimized the ML architecture of the app is. Consider the main components that are part of computing resources and influence the operation of mobile apps:

Table 1 - Main components that are part of computing resources and influence the operation of mobile apps

Elements	Characteristics and explanation
Central processing unit (CPU) as a basic computing element	The CPU is responsible for generating general computational loads and maintaining the logic flow of the app. In the context of mobile app personalization, it: intermediates between applications and RAM, exchanges information with the other chips (GPU, NPU); performs ML computations with low bandwidth; performs automated data preprocessing (normalization, filtration, aggregation); distributes computational load among chips, ensures operation of models in cases of unavailable specialized blocks. Limitations: low bandwidth, inefficient parallel computing, raised power consumption during model inference.

Graphics processing unit (GPU) for parallel computing	GPUs provide an exponentially larger number of computational cores and allow to speed up parallel computing in chips of mobile devices. In mobile devices, GPUs are used for: processing multimedia data (images, videos), as they are the most frequent sources of data for AI personalization; performing specific ML computations optimized for visuals (matrix multiplication). Nevertheless, GPU demands a lot of power and, thus, can be inefficient for performing long computing pipelines.
Neural processor (NP) for accelerating neural network computations	The NPU is optimized for: performance and DL operations; performing model inference; executing local personalized content without contribution of cloud resources; reducing power consumption. The neural processor allows quickly producing such personalization tasks as research of recommendations, analysis of user activity, classifying content, real, time generation of user, specific settings.
A data processing accelerator (DPA) combining central processing unit (CPU) and graphics processing unit (GPU) functions on a single chip for joint computing and mutual acceleration.	DPA (Deep Learning Processing Accelerator, formerly known as a hardware accelerator) is a high performance neural processing unit integrated into SoC of the system, combining the capabilities of CPU, GPU, and the AI specialist cores. A DPA by MediaTek allows to: accelerate deep learning algorithms; work with mixed architectures combining CNNs, RNNs, and transformer blocks; reduce power consumption by 4 to 10 times (compared with CPU and GPU); use computational precisions of 16, bit floating point (FP16),8, bit integer (INT8) and bit integer (INT4) (on 7, 8 times, cheaper than FP32). The same task in the GPU or CPU will be 2, 4 times longer and consume more power. DPA also supports MediaTek optimizations and proprietary improvements over Neural Engine and GPU. The main possible opportunities for using DPAs in smartphone personalization: processing data and personalized models without Internet; generating local recommendations for every day without model transfer to the server, respective contextual adjustments of the interface (according to user activity and preferences); recognition of objects, gestures, speech signals as parts of interactive environments. A DPA, enabled smartphone accelerates on, device personalization and minimizes risks of user data disclosure.
RAM, bandwidth, and permanent memory on the device	Personalized models have to work smoothly with the user data stream and to be supported with RAM (to keep the model and user, specific data in memory, cache recommendations and interaction setting history). Also, the performance of large predictive models practically depends on memory and bandwidth capabilities of the mobile device. Persistent memory (the device storage) is used for long, term data, models, and ML cache storage.

Source: compiled by the author

Modern mobile devices are often powered by specialized hardware accelerators capable of running AI algorithms directly on the device. MediaTek incorporates NPUs (Neural Processing Unit) and DPAs (Deep Learning Processing Accelerators) into their chipsets, enabling faster and more efficient processing of ML and DL models. The NPU (Neural Processing Unit) is a multi-core processor, created to provide acceleration to DL and computer vision tasks in the lowest latencies and with the lowest energy consumption possible. The term DPA (Deep Learning Processing Accelerators) previously used term to denote a similar hardware accelerator is considered nowadays part of modern MediaTek chipsets' NPU or an older generation one.

Using an NPU, an on-device personalization, predictive recommendations and context aware operations are made possible on a mobile application, without burdening the CPU or GPU with complex tasks. Such an approach not only makes intelligent features fast but also minimizes power usage and ensures privacy. The NPU also supports performance scaling in newer chipsets by MediaTek (Dimensity 9000 and above), allowing computing resources to adapt to the complexity of an AI model to ensure efficiency, a trade off between predictive accuracy and efficient compute usage. Consequently, integration of the NPU and DPA on mobile platforms contributes to overcoming one of the obstacles for personalized mobile applications - insufficient compute resource - by augmenting the abilities of adaptive and context sensitive on-device AI.

The Moto AI, Xiaomi Hyper AI, Apple Intelligence and Galaxy AI ecosystems driven by AI are connected to the use of smart devices. These systems connect various devices and services together and function with its specific AI platforms so to provide performance enhancement, integration and user experience, voice assistance, AI camera and automatic data processing, cloud service combination.

Neural processors exist on the device as part of mobile ecosystem, for example Qualcomm Hexagon DSP and Apple Neural Engine, are designed to make the computation of AI on the device, and they provide efficient and low-latency operations for intensive machine learning applications, image processing, and voice recognition tasks.

Conclusions

In the paper, we provided a discussion of machine learning based, deep learning based, NLP based, real time analytics based and context aware personalization techniques are discussed and proved that classical recommendation techniques combined with modern neural networks gives the highest accuracy and the best adaptability of personalized services.

It was shown that one of the major obstacles in deploying AI on mobile application is resources constraint of mobile devices, variety of OS, user dynamism and guarantee of privacy data. It was concluded that personalization can be implemented effectively by performing model optimization, light architecture design and contemporary privacy preservation techniques.

The study identified that confidential personalization frameworks, more specifically federated learning and differential privacy, serve as cornerstones in the construction of secured personalization systems enabling training of models locally without exposing the user information to the server.

Having analyzed the hardware platforms we have found out that specialized neural hardware processors such as the neural processing units (NPUs) and deep learning processing accelerators (DPAs) are fundamental to support effective execution of the AI algorithms on the mobile devices to achieve low power consumption, high performance and real-time execution of the personalized models on the device itself, without utilizing any cloud.

Furthermore, it has been observed that new generation mobile AI ecosystem including Galaxy AI, Apple Intelligence, Xiaomi Hyper AI and Moto AI is creating a new tier of software/hardware integration, which greatly enhances personalization and contextuality of personalization.

Overall, the experimental results show that, although the personalized approaches are well-developed, no uniform architecture is presented capable of integrating performance, privacy, inclusivity and adaptivity to the context all together. That can dictate the future research direction to build scalable, energy-efficient and context-aware AI on the mobile devices.

Future research may investigate thoroughly in the performance of neural processors when implementing the personal models of different complexities and energy efficient operation during local inference in mobile devices, as well as the adaptation of transformed model in resource-constraint mobile environment.

References:

1. Bronzin, T., Prole, B., Stipić, A., & Pap, K. (2021, September). Artificial Intelligence (AI) brings enhanced personalized user experience. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, 1070-1075. https://doi.org/10.23919/MIPRO52101.2021.9596938
2. Okonkwo, C. (2024). Assessment of user experience (UX) design trends in mobile applications. *Journal of Technology and Systems*, *6*(5), 29-41. URL: https://carijournals.org/journals/JTS/article/view/2147/2535
3. Nama, P. (2023). AI-Powered Mobile Applications: Revolutionizing User Interaction Through Intelligent Features and Context-Aware Services. *Journal of Emerging Technologies and Innovative Research*, *10*(01), g611-g620. URL: https://www.researchgate.net/publication/385207252_AI-Powered_Mobile_Applications_Revolutionizing_User_Interaction_Through_Intelligent_Features_and_Context-Aware_Services
4. Nishant Agrawal (2025). How is AI revolutionizing mobile app development. *Competenza Innovare Think innovative. Deliver excellence*. URL: https://competenza.ae/blog/ai-in-mobile-app-development-benefits-guide/
5. Gupta K. (2025) AI-POWERED ACCESSIBILITY IN ANDROID APPLICATIONS: ADVANCING DIGITAL INCLUSION THROUGH MACHINE LEARNING. *International Journal of Information Technology and Management Information Systems (IJITMIS), 16*(1), 659-671. https://doi.org/10.34218/IJITMIS_16_01_047
6. Manikanthan, S. V., Padmapriya, T., Hussain, A., & Thamizharasi, E. (2020). Artificial intelligence techniques for enhancing smartphone application development on mobile computing. https://doi.org/10.3991/ijim.v14i17.16569
7. Sarker, I. H., Khan, A. I., Abushark, Y. B., & Alsolami, F. (2021). Mobile expert system: exploring context-aware machine learning rules for personalized decision-making in mobile applications. *Symmetry*, *13*(10), 1-10. https://doi.org/10.3390/sym13101975
8. Sefa-Yeboah, S. M., Osei Annor, K., Koomson, V. J., Saalia, F. K., Steiner-Asiedu, M., & Mills, G. A. (2021). Development of a mobile application platform for self-management of obesity using artificial intelligence techniques. *International Journal of Telemedicine and Applications*, *2021*(1), 1-16. https://doi.org/10.1155/2021/6624057
9. Nama, N. P. (2021). Enhancing user experience in mobile applications through AI-driven personalization and adaptive learning algorithms. *World Journal of Advanced Engineering Technology and Sciences*, *3*(2), 083-094. https://doi.org/10.30574/wjaets.2021.3.2.0064
10. Nwanna, M., Offiong, E., Ogidan, T., Fagbohun, O., Ifaturoti, A., & Fasogbon, O. (2025). AI-driven personalisation: Transforming user experience across mobile applications. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *3*(1), 1930-1937. https://doi.org/10.51219/JAIMLD/maxwell-nwanna/425

11. Kumar, P., Kumar, N., & Jain, A. (2023). Enhancing Mobile App Performance with AI-Driven Optimization Techniques. *International Journal of Computer Technology and Electronics Communication*, *6*(6), 7893-7903. https://doi.org/10.15680/IJCTECE.2023.0606006
12. Li, Y., Dang, X., Tian, H., Sun, T., Wang, Z., Ma, L., ... & Bissyandé, T. F. (2025). An empirical study of AI techniques in mobile applications. *Journal of Systems and Software*, *219*, 1-23. https://doi.org/10.1016/j.jss.2024.112233
13. Madhavan, N. (2024). The Role of Artificial Intelligence in Enhancing Mobile App Accessibility. *International Journal of Advanced Research and Emerging Trends (JARET*), 1(2), 1-12. URL: https://www.researchgate.net/publication/383115711_The_Role_of_Artificial_Intelligence_in_Enhancing_Mobile_App_Accessibility_Inclusive_User_Experience_Digital_Inclusion
14. Goyal, V., & Singh, S. (2025). AI-Powered Accessibility in Mobile Applications: Advancing Inclusive Design through Machine Learning Innovations. *International Journal of Scientific Development and Research*, *10*(2), a263-a279. URL: https://www.researchgate.net/publication/391526472_AI-Powered_Accessibility_in_Mobile_Applications_Advancing_Inclusive_Design_through_Machine_Learning_Innovations
15. Namoun, A., Alrehaili, A., Nisa, Z. U., Almoamari, H., & Tufail, A. (2024). Predicting the usability of mobile applications using AI tools: the rise of large user interface models, opportunities, and challenges. *Procedia Computer Science*, *238*, 671-682. https://doi.org/10.1016/j.procs.2024.06.076
16. Artificial Intelligence App Development: Transforming Mobile Apps. (2024) *Stfalcon*. URL: https://stfalcon.com/en/blog/post/the-impact-of-artificial-intelligence-on-mobile-app-development
17. Intelegain Technologies (2025). AI-driven Personalization: How Apps Are Becoming Smarter and More Intuitive. URL: https://www.intelegain.com/ai-driven-personalization-how-apps-are-becoming-smarter-and-more-intuitive/
18. Amin, M. (2024). AI-Driven Configuration Management for IP Network Devices Using XML and Machine Learning. *configurations*, *18*, 1-12. https://doi.org/10.20944/preprints202410.2197.v1
19. Collaborative Filtering: Types, Pros & Challenges. *Bot Penguin.* URL: https://botpenguin.com/glossary/collaborative-filtering
20. Natural Language Processing Applications to Watch Now. *Intellspot.* URL: https://www.intellspot.com/natural-language-processing-applications/