# OPTIMIZING MACHINE LEARNING ARCHITECTURES FOR MOBILE DEVICES: MODELS, PERFORMANCE, AND PRIVACY

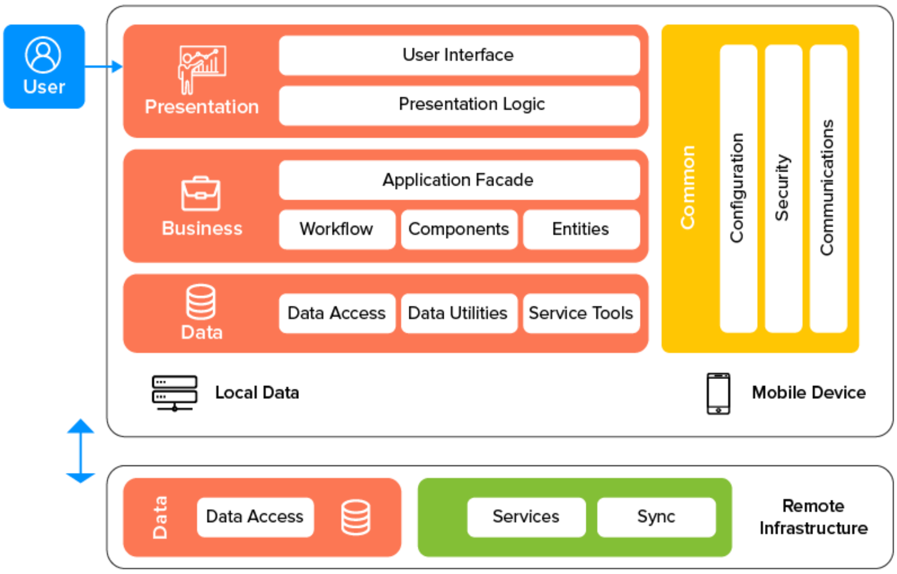
**Adrian Runceanu,** *"Constantin Brâncuşi" University of Târgu Jiu, ROMANIA*
**Mihaela-Ana Runceanu,** *"Ecaterina Teodoroiu" High College, Târgu Jiu, ROMANIA*

**ABSTRACT:** Modern mobile applications rely heavily on Machine Learning technology to provide image recognition and natural language processing capabilities, as well as tailored recommendation features. Implementing Machine Learning systems in mobile settings requires informed architectural choices that balance performance needs with device resource constraints and safeguard user data confidentiality. This study examines two primary execution techniques, encompassing device-based inference and cloud-based processing, to assess their impact on system speed and accuracy, network demands, and user information security. The article examines three optimisation methods that enable mobile devices to run machine learning models efficiently. We investigate battery usage and thermal management issues alongside new privacy-enhancing solutions that comprise device-based processing and distributed learning systems. Studies have shown that streamlined system configurations lead to enhanced user satisfaction by providing faster performance and safeguarding individual data. Studies have found that well-designed systems enhance users' experiences by providing enhanced performance and safeguarding sensitive data.

**KEY WORDS:** Mobile Machine Learning, Cloud-Based ML, Cloud Offloading, Model Optimization, Mobile Privacy & Security.

## 1. INTRODUCTION

Implementing machine learning models into mobile applications necessitates tailored architectural solutions that can effectively manage the limited resources of mobile devices, encompassing memory constraints, power consumption, and processing speed limitations. The system necessitates a strategic approach to identify model execution sites between local devices and cloud servers while



managing data and preserving user privacy protection.

Figure 1: The most popular multilayer architecture is the three-layer architecture. [16]

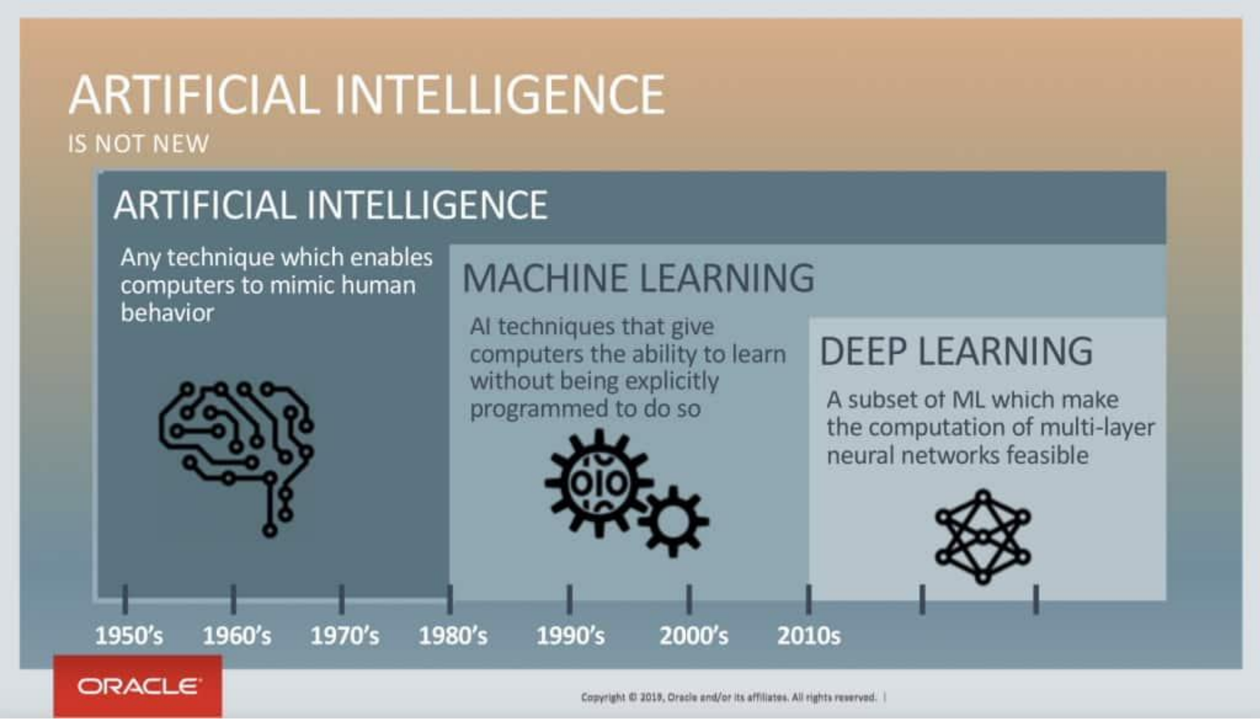
## 2. ON-DEVICE MODELS VS. CLOUD-BASED MODELS

Developers working on mobile applications must choose between running machine learning models directly on a device or sending them to cloud servers for processing.

The choice between local and cloud-based model execution influences application performance, system resource utilisation, and the level of data security offered to users. The application can function normally without internet connectivity due to local model processing. Storing personal data on mobile devices allows users to have more control over their information and enjoy enhanced privacy levels. The system is subject to performance limitations due to its reliance on the device's hardware capabilities, such as RAM capacity and processing velocity [1].Data processing typically takes place on sophisticated servers, which have more advanced computational capabilities than typical local devices. The system functions optimally for managing complex operations and large dataset access needs. The solution offers superior performance, but users need to ensure persistent internet connectivity, and their personal data is exposed during server-to-server data transfers [3].

Table 1 below illustrates the key distinctions between machine learning models that run directly on user devices (on-device) and those that operate on external servers (cloud-based). The information synthesized is from official documentation for Machine Learning (ML) Kit, Oracle, and IBM Cloud. The table illustrates key considerations that developers need to take into account when deciding between on-device and cloud-based models for their mobile applications. The success of mobile applications that incorporate machine learning features is determined by the selection process, which is dependent on these fundamental elements.

The Google board functions as a demonstration of in-device model deployment because it offers real-time word recommendations without transmitting user data to remote servers thereby safeguarding user confidentiality. Google Translates cloud-based models provide translation services online, and they also power Google Photos' object recognition, requiring access to robust servers and large databases. The models' accuracy is improved, but users must maintain an active internet connection for their operation [4], [22], [23].



ARTIFICIAL INTELLIGENCE
IS NOT NEW

ARTIFICIAL INTELLIGENCE
Any technique which enables computers to mimic human behavior

MACHINE LEARNING
AI techniques that give computers the ability to learn without being explicitly programmed to do so

DEEP LEARNING
A subset of ML which make the computation of multi-layer neural networks feasible

Copyright © 2019, Oracle and/or its affiliates. All rights reserved. |

Figure 2: Differences between AI, Machine Learning and Deep Learning. [17]

Table 1. Comparison between on-device and cloud-based machine learning models

| Characteristic | On-Device Models | Cloud-Based Models |
|---|---|---|
| Internet connection required | No | Yes |
| Data privacy | High (data never leaves the device) | Low (data is transmitted online) |
| Response speed | Instant (no network latency) | Slower (network-dependent) |
| Model complexity | Limited by device hardware | Higher, no local constraints |
| Resource consumption | Higher (local processing) | Lower on the device |
| Examples | Offline translation, local face recognition | Alexa, Google Assistant, Siri (partially) |

## 3. DATA PROCESSING AND MODEL OPTIMIZATION FOR MOBILE DEVICES

Optimization of machine learning models and data processing for mobile devices for optimal performance, mobile applications that employ artificial intelligence must have machine learning models that adapt to the hardware limitations of smartphones and tablets. Mobile devices are limited by their processing power, memory capacity, and power consumption, necessitating the optimization of resource usage. Reducing the size of ML models is necessary to enable them to operate quickly and with minimal energy consumption. The system should handle all data processing tasks, encompassing images, text, and audio, directly on the device, as this method provides rapid responses and protects user privacy [5], [2]. Mobile applications must be optimized for performance given the limited resources available on mobile devices. Image-processing applications speed up their operations by reducing resolution prior to beginning their analysis. The natural language processing system performs optimally when users input simple sentences, and it processes smaller amounts of data. The natural language processing system which includes automatic translation and chatbots functions best when users enter basic sentences and the system handles reduced amounts of data. The server-based versions of machine learning models need to be significantly downsized and improved in terms of efficiency in order to be adapted for mobile phone use. Running complex models with millions of parameters on a smartphone becomes impractical. Developers implement three primary optimization techniques to improve model performance, which comprise model component removal, operation streamlining, and architectural simplification. Optimizing applications leads to enhanced performance while reducing resource consumption [2], [20], [21]. Implementing processing directly on mobile devices provides a highly effective solution for mobile machine learning applications. The system allows for rapid data processing while eliminating the need for a network and protecting user data from unauthorized access. Prior to analysis, the model compresses and simplifies images and optimizes text preprocessing to expedite operations [5].

Developers must implement particular techniques when optimizing ML models for mobile devices as outlined in Table 2. TensorFlow Lite and ML Kit optimization tools empower developers to deploy optimization techniques and run models with high performance on mobile devices [2].

Personal user data is processed locally on the device, safeguarding user privacy from potential external server breaches. The demands for significant computational power and access to large datasets necessitate the use of cloud-based solutions, which in turn introduces reliance on network systems [4], [24], [25]. The table below displays actual mobile applications which use model optimization methods to achieve efficient operation on handheld devices.

| Technique | Description |
|---|---|
| **Pruning** | removing non-essential parts of the model to reduce its size and improve speed. |
| **Quantization** | reducing numerical precision to lower memory usage and increase processing speed. |
| **Knowledge distillation** | creating a smaller model that learns to replicate the behavior of a larger, more complex model while maintaining comparable performance. |

Table 2. Techniques for optimize ML models

| Mobile application | Description |
|---|---|
| **Google Lens** | is a popular mobile application that uses artificial intelligence to recognize objects, text, and scenes from the real world. To run quickly and efficiently directly on the phone without constant internet access, its underlying models are optimized through techniques such as quantization and pruning, which make them smaller and faster. |
| **Snapchat** | uses advanced image processing methods and augmented reality (AR) technology to apply filters and effects to users' faces. These functionalities rely on knowledge distillation, in which large models running on servers are used to train smaller models capable of operating directly on mobile devices. This approach not only reduces computational resource usage but also lowers energy consumption. Additionally, Snapchat applies quantization techniques to increase image-processing speed and reduce memory usage [6]. |
| **Google Translate** | employs advanced machine learning techniques to translate text between languages. The core translation models are initially trained on powerful servers, then simplified through quantization and pruning to make them suitable for mobile use, including offline mode. To further accelerate translation, the application can reduce sentence complexity before processing, enabling a faster and more efficient user experience [7]. |
| **Spotify** | implements personalized recommendation algorithms based on machine learning. The models analyzing user preferences are optimized through knowledge distillation, enabling the creation of compact versions of complex server-side models that can run efficiently on mobile devices. The application also employs pruning to remove irrelevant parameters, reducing memory usage and improving response times [8]. |
| **Apple Face ID** | uses machine learning techniques for facial authentication. The facial recognition model is optimized to run efficiently on mobile hardware, ensuring a fast and seamless unlocking experience. Furthermore, all processing occurs locally on the device, guaranteeing the protection of personal biometric data [9]. |
| **Instagram** | employs machine learning techniques to personalize user feeds and provide content recommendations. The recommendation models are optimized for mobile performance, enabling fast processing and smooth user interaction. The application uses knowledge distillation to generate compact versions of the models that can run efficiently without excessive resource consumption. These ML optimization techniques contribute to the improved performance of popular mobile applications, ensuring responsiveness and efficiency even on devices with limited capabilities. Moreover, local data processing helps reduce memory usage and enhances user privacy protection [10], [11]. |

Table 3. Mobile applications that use model optimization techniques in their development process.

## 4. PERFORMANCE AND RESOURCE CONSUMPTION CONSIDERATIONS

Performance and resource utilization factors to consider optimizing performance and managing resources is crucial for the implementation of machine learning in mobile apps. Mobile devices run on limited processing power and limited memory, and their batteries have relatively short lifespans. Machine learning applications must run at peak efficiency as they are deployed on mobile devices with restricted system resources. Smartphone performance can be improved by using specific optimization techniques that allow machine learning models to operate efficiently without affecting battery life or device speed.

Reducing a model's size constitutes a basic optimization method. Despite the complexity of machine learning models, mobile devices can operate effectively without the full model architecture. Model size reduction is achieved through two methods: one involves deleting unnecessary model components, and the other involves reducing numerical values. This optimization enables the application to run more efficiently and at a faster speed. The application becomes more efficient and runs at increased speed because of this optimization.

## 5. CONCLUSION

The study examined the functioning of machine learning models on mobile devices, focusing on their underlying architectures and methods to boost performance and meet security demands. The speed of a mobile application hinges on user data protection since users must choose between processing models locally or transmitting data to cloud servers for analysis. On-device model execution offers rapid processing and robust data security, but it is constrained by hardware limitations, whereas cloud-based processing provides substantial computing power, though it necessitates a stable network and involves a trade-off in user privacy. Three optimization techniques enable efficient performance by reducing model complexity and computational needs for mobile ML operations. Mid-range devices can deliver real-time performance by using hardware accelerators in conjunction with on-device data preprocessing and these optimization techniques. Federated learning with secure aggregation enables the development of improved models by protecting user confidentiality and meeting all relevant regulatory standards [23], [26].

Achieving a balance between system accuracy, operational speed, and user trust requires meticulous planning for the development of Mobile ML systems. The research will concentrate on developing hybrid execution systems, adaptive optimization techniques, and privacy protection frameworks to support intricate mobile machine learning applications, given the ongoing advancements in mobile technology and increasing stringent privacy requirements.

The development of Mobile ML systems needs careful planning to achieve the best possible combination of system accuracy and operational speed and user trust. The research will focus on creating hybrid execution systems and adaptive optimization techniques and privacy protection frameworks which will support complex mobile ML applications because mobile technology continues to advance and privacy standards become more stringent.

## REFERENCES

[1] Google Developers, „ML Kit", 2024. Accessed online at: https://developers.google.com/ml-kit

[2] TensorFlow. (2023). Optimize models for mobile and edge with TensorFlow Lite. Accessed online at: https://www.tensorflow.org/lite/performance/model_optimization

[3] Oracle (2024) What is a cloud database? Oracle Romania. Accessed online at: https://www.oracle.com/ro/database/what-is-a-cloud-database/

[4] IBM (2023). What is acloud computing?IBM Cloud Education. Accessed

online at:
https://www.ibm.com/cloud/learn/cloud-computing

[5] IBM (2023). What is a cloud database? Oracle Romania. Accessed online at:
https://www.ibm.com/cloud/learn/machine-learning

[6] 5 Ways Snapchat Uses Artificial Intelligence and Machine Learning. Accessed online at:
https://daanishbhatti.medium.com/5-ways-snapchat-uses-artificial-intelligence-and-machine-learning-a885d29eed66

[7] Inteligenţa Artificială a transformat Google Translate şi are capacitatea să reinventeze IT-ul. Accessed online at:
https://www.descopera.ro/lumea-digitala/16028897-inteligenta-artificiala-a-transformat-google-translate-si-are-capacitatea-sa-reinventeze-it-ul

[8] How Spotify uses Machine Learning? Accessed online at:
https://www.projectpro.io/article/how-spotify-uses-machine-learning/687

[9] About Face ID advanced technology. Accessed online at:
https://support.apple.com/ro-ro/102381

[10] Instagram explains how it uses AI to choose content for your Explore tab, Accessed online at:
https://www.theverge.com/2019/11/25/20977734/instagram-ai-algorithm-explore-tab-machine-learning-method

[11] How Instagram ads use machine learning. Accessed online at:
https://help.instagram.com/119516847899397/?helpref=related_articles

[16] https://os-system.com/blog/mobile-app-architecture-how-to-design-it/

[17]
https://blogs.oracle.com/bigdata/post/whatx27s-the-difference-between-ai-machine-learning-and-deep-learning

[18] Grofu Florin, "Fast Data Acquisition System", Revista de Fiabilitate și Durabilitate Nr 1/2022 Editura "Academica Brâncuşi", Târgu Jiu, ISSN 1844 – 640X Pg. 49-54

[20] Florin Grofu, Constantin Cercel, "Experimental module for training with network analyzer", 12th International Multidisciplinary Scientific GeoConference & EXPO SGEM2012, Vol III- section Informatics, ISSN: 1314-2704, pg 111-117

[21] Gheorghe GÎLCĂ, Nicu George BÎZDOACĂ, Cătălin LUPU, "Classification algorithms of facial expressions by using the feedforward neural networks", National Scientific Conference with international participation "CONFERENG 2015", Tg-Jiu, November 13-14 2015, Nr. 4, pp. 80-85, ISSN 1842-4856, Annals of the „Constantin Brâncuşi" University from Tg. Jiu, Engineering Series

[22] Gheorghe GÎLCĂ, Nicu George BÎZDOACĂ, Cătălin LUPU, "Artificial Vision Algorithms based on Fractals for the Images Description", National Scientific Conference with international participation "CONFERENG 2015", Tg-Jiu, November 13-14 2015, Nr. 4, pp. 86-90, ISSN 1842-4856, Annals of the „Constantin Brâncuşi" University from Tg. Jiu, Engineering Series

[23] Ilie Borcosi, Corina Ana Borcosi, "Customize SAAS applications through microservices", Annals of the „Constantin Brancusi" University of Targu Jiu, Engineering Series, No. 2/2022, pag. 86-90.

[24] Ilie Borcosi, Daniela-Lavinia Nebunu, Nicolae Antonie, "The modeling logic circuit using finite automata theory", International Conference on Industrial Electronics Technology & Automation (IETA 11) in International Joint Conferences on Computer, Information and Systems Sciences, and Engineering (CISSE11), December 3– 12, 2011

[25] Ionescu, M., Nicu-George BÎZDOACĂ, "The control of a biomimetic structure using the pid algorithm", National Scientific Conference with international participation "CONFERENG 2017", Tîrgu-Jiu, December 08-09, No. 4, Year 2017, pp. 113-118, ISSN 1842-4856, Annals of the „Constantin Brancusi" University of Targu Jiu, Engineering Series.

[26] Ionescu, M., Vilan, C., Dinca, A. "Control architecture the for a biomimetic structure study", 13th International Multidisciplinary Scientific Geoconference SGEM 2013, 16-22 June, 2013, Albena, Bulgaria, volume I, ISBN 978-954-91818-9-0, ISSN 1314-2704, pp.45-52, WOS:000349067200006