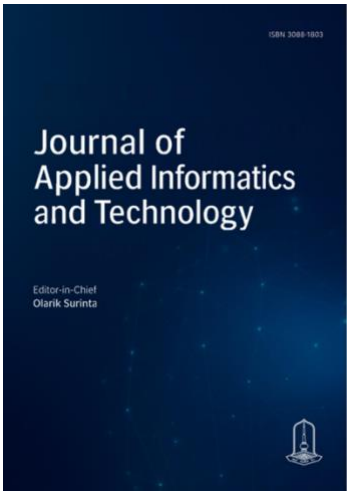
Pre-proof



A Unified OCR-LSTM Pipeline for Personalized Running Performance Analysis and Forecasting

Kawisara Prasitnui, Natthaphon Reungwaraphit, and Sunee Pongpinigpinyo

Article ID: 266006

Received: 2 February 2026
Revised: 20 August 2026
Accepted: 25 August 2026

Please cite this article as: Prasitnui, K. (2027). A Unified OCR-LSTM Pipeline for Personalized Running Performance Analysis and Forecasting, *Journal of Applied Informatics and Technology*, https://ph01.tci-thaijo.org/index.php/jait/article/view/266006

This is a PDF of an article that has been peer-accepted and formatted for early visibility, including the addition of a cover page and metadata. This version will undergo final copyediting, typesetting, and proof review before publication as the definitive Version of Record. During this production process, minor changes or errors may be identified and corrected. Please note that the journal's sharing and copyright policies for published articles apply to this advance version.

© 2026 Published by Faculty of Informatics, Mahasarakham University.

RESEARCH ARTICLE

A Unified OCR-LSTM Pipeline for Personalized Running Performance Analysis and Forecasting

Kawisara Prasitnui[1], Natthaphon Reungwaraphit[1], Sunee Pongpinigpinyo[1,*]

[1]Department of Computing, Faculty of Science, Silpakorn University, Nakhon Pathom 73000, Thailand

* Corresponding author: Sunee Pongpinigpinyo (pongpinigpinyo_s@su.ac.th)

Article Information

Article History:

Keywords:
Deep Learning
LSTM
Mobile Application
Optical Character Recognition
Running Analytics
Time Series Forecasting

Abstract

The analysis of running performance is often challenged by data fragmentation across various tracking applications. This study presents a novel mobile application that automates data collection and provides personalized performance forecasting by integrating Optical Character Recognition (OCR) with a Long Short-Term Memory (LSTM) deep learning model. To address data fragmentation, the system allows users to upload running performance summary images, from which an EasyOCR-based module extracts key metrics such as distance and duration. These time-series data are subsequently processed by an LSTM network for two primary tasks: forecasting future running performance over a 30-day horizon and classifying daily activity (run vs. non-run). Experimental results demonstrated the effectiveness of the OCR component, which achieved a low average Word Error Rate (WER) of 10.3%, with 70% of test images recognized perfectly. The LSTM forecasting model yielded a Mean Absolute Error (MAE) of 5.3417 km for distance and 21.1524 minutes for duration. Meanwhile, the binary classification model achieved an accuracy of 68%, with balanced F1-scores of 0.69 (non-running) and 0.67 (running). Furthermore, usability testing confirmed a 100% task completion rate, with automated extraction averaging 18.02 seconds and re-upload times dropping to 8.53 seconds, demonstrating high learnability. These findings validate the feasibility of a unified OCR-LSTM pipeline for running analytics.

1. Introduction

Running is one of the most popular forms of physical activity globally, offering significant health benefits, including improved cardiovascular health, weight management, and mental well-being. The World Health Organization (WHO) recommends regular physical activity to reduce the risk of noncommunicable diseases (World Health Organization, 2024). With the advent of wearable technology and fitness tracking applications, runners now have access to vast amounts of data regarding their performance. However, a significant challenge of data fragmentation persists. Runners often switch between different applications or devices, leading to scattered records that are difficult to consolidate for long-term analysis. Furthermore, while many existing applications provide historical data, few offer automated digitization from running performance summary images or personalized predictive analytics to guide future training.

Current solutions for data digitization, such as text readers for the visually impaired, utilize Optical Character Recognition (OCR) technology (Manikandan et al., 2017). While machine learning has shown great potential in health analytics, there is still a lack of integrated systems that combine automated data collection from images with deep learning-based time series forecasting.

To overcome these challenges, this research proposes "Smart Run," a mobile application developed to address these issues. The application is designed to automate the digitization of running statistics from summary images using OCR, ensuring data integrity through verification logic, and to forecast future running performance (distance and duration) and activity patterns using Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997). By leveraging a large-scale dataset of long-distance running records, the system provides personalized trend analysis and event recommendations, thereby enhancing the user's ability to plan and improve their physical activity effectively.

JIT Pre-Proof

Table 1. Summary of related works in health and activity prediction.

Related work	Method used	Result
A Study on Calories Burnt Prediction using Machine Learning (Panwar et al., 2023)	Linear Regression, XG boost Regressor, and Random Forest	Best Mean Absolute Error (MAE) = 1.48 on XG boost Regressor
Machine Learning modeling for predicting adherence to physical activity guideline (Choe et al., 2025)	Logistic Regression, Random Forest, Gradient Boosting (XGBoost), and SVM	Accuracy = 70.5%, Area Under Curve (AUC) = 0.542, F1-score = 0.819 (best from decision tree)
Comparison of Machine Learning Methods for Calories Burn Prediction (Tan Jing Sheng et al., 2024)	Linear Regression, XG boost Regressor, Random Forest, and LightBGM Regressor	Best Mean Absolute Error (MAE) = 1.27 on LightBGM Regressor

2. Related work

This section discusses research efforts related to physiological and behavioral prediction involving machine learning techniques. The first sub-section of this article discusses published research works in the calories burnt prediction domain. The following subsection summarizes other predictive algorithms that have been successfully applied to related health domains, such as physical activity adherence analysis. Several papers have been written on the topic of estimating energy expenditure and health-related behaviors using regression and classification models. A detailed summary is described in Table 1.

3. Materials and Methods

3.1 System architecture

The proposed "Smart Run" application operates on a client-server architecture designed to balance user responsiveness with computational efficiency, as shown in Figure 1. The mobile client was developed using the Flutter framework (Google, 2024), enabling cross-platform compatibility for Android devices. The client handles user interactions, images, and data visualization. The server-side backend was implemented using Python with the Flask framework (Pallets, 2024), serving as an API gateway for processing requests. Firebase Firestore (Google Developers, 2024), a NoSQL cloud database, is utilized for scalable data storage, managing user profiles, running event logs, and authentication via Google and Facebook OAuth 2.0 standards (IETF, 2012). The system's core analytical engine integrates two key subsystems: an Optical Character Recognition (OCR) for the automated digitization of running statistics and a Deep Learning dedicated to time-series forecasting.

3.2 Automated data extraction using Optical Character Recognition

The system addresses data fragmentation by utilizing EasyOCR (Jaided AI, 2020), which integrates CRAFT (Baek et al., 2019) for text detection and CRNN (Shi et al., 2017) for recognition, to process fitness summary images. Beyond alphanumeric extraction, the engine retrieves spatial coordinates to dynamically render bounding boxes on the mobile interface, providing a visual mechanism for user verification. The extracted text is parsed into structured key-value pairs: Date, Distance, Duration, Pace, and Heart Rate. Finally, to ensure data integrity, a heuristic verification algorithm validates mathematical consistency

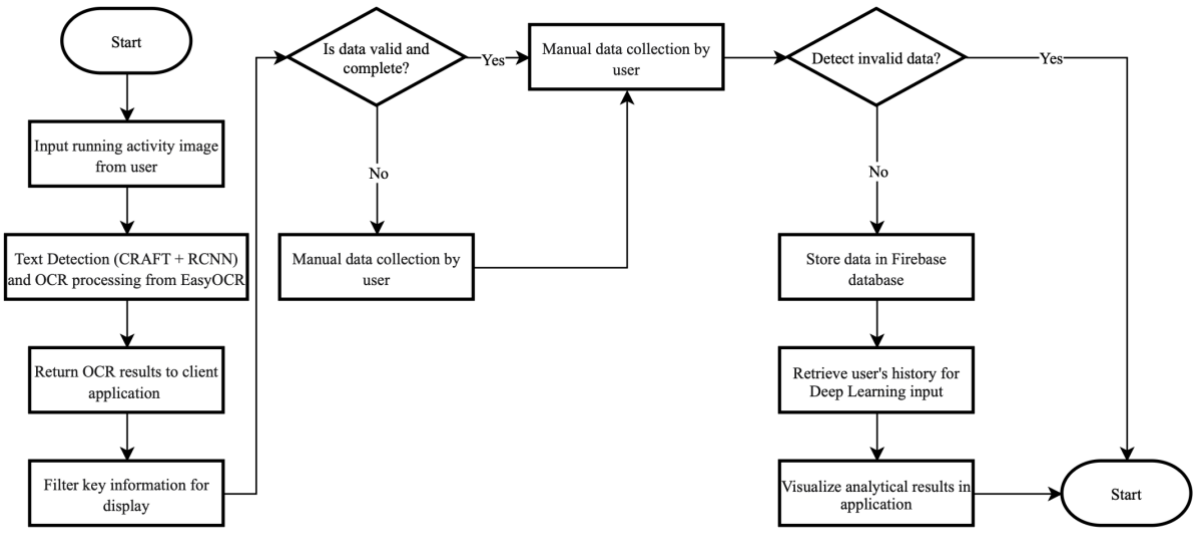


Figure 1. System architecture of Smart Run application.

(e.g., Duration = Distance × Pace) prior to user confirmation.

3.3 OCR Performance Evaluation

The accuracy of the OCR system was evaluated using the Word Error Rate (WER) metric (Morris et al., 2004), as defined in Equation (1), to quantify the reliability of the extraction process.

$$WER = \frac{S+D+I}{N} \quad (1)$$

where S, D, and I represent the number of substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference text.

3.4 Datasets and Preprocessing

The study utilized the "Long-distance running dataset" from Kaggle (Mexwell, 2023), comprising approximately 10.7 million records from 36,412 anonymized athletes. The preprocessing pipeline began with filtering data based on activity frequency to enhance model reliability. This process involved excluding the bottom 10th percentile for regression models to capture consistent training behaviors and applying a stricter 40th percentile threshold for classification to clearly distinguish activity patterns. Concurrently, a binary running feature was added, yielding a balanced dataset of 9,775,995 running (Class 1) and 10,162,965 non-running (Class 0) instances. Subsequently, the dataset was partitioned by unique athlete IDs into training (80%), validation (10%), and testing (10%) sets. Following the split, feature engineering was applied where numerical features were normalized using a Min-Max Scaler, and categorical variables were one-hot encoded. Finally, 30-day sliding window sequences were generated separately for the training and validation sets to train the model, and for the test set to evaluate performance.

3.5 Deep Learning Model Architecture

A Bidirectional LSTM (Schuster & Paliwal, 1997) architecture was implemented using TensorFlow to capture long-term dependencies in sequential data. The system comprises three models: the Distance and Duration models for regression tasks shown in Figure 2, and the IsRunning model for binary classification to predict daily activity probability shown in Figure 3.

The architecture comprises three stacked Bidirectional LSTM layers (256, 128, and 64 units), interleaved with Dropout layers (rate 0.2) (Srivastava et al., 2014) to prevent overfitting. The final Dense output layer utilizes ReLU activation for regression models and Sigmoid for the classification task.

The models were trained using the Adam optimizer. To handle outliers inherent in user-generated fitness data, the regression models utilized the Huber Loss function (Huber, 1992), as defined in Equation (2).

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & if\ |a| \le \delta \\ \delta \cdot (|a| - \frac{1}{2}\delta) & otherwise \end{cases} \quad (2)$$

where a represents the residual $(y - \hat{y})$ and δ the transition threshold. The classification model utilized Binary Cross-Entropy Loss. To prevent overfitting, Early Stopping halted training upon validation loss stagnation, automatically restoring optimal weights. Finally, an Integrated Forecasting approach modulated regression

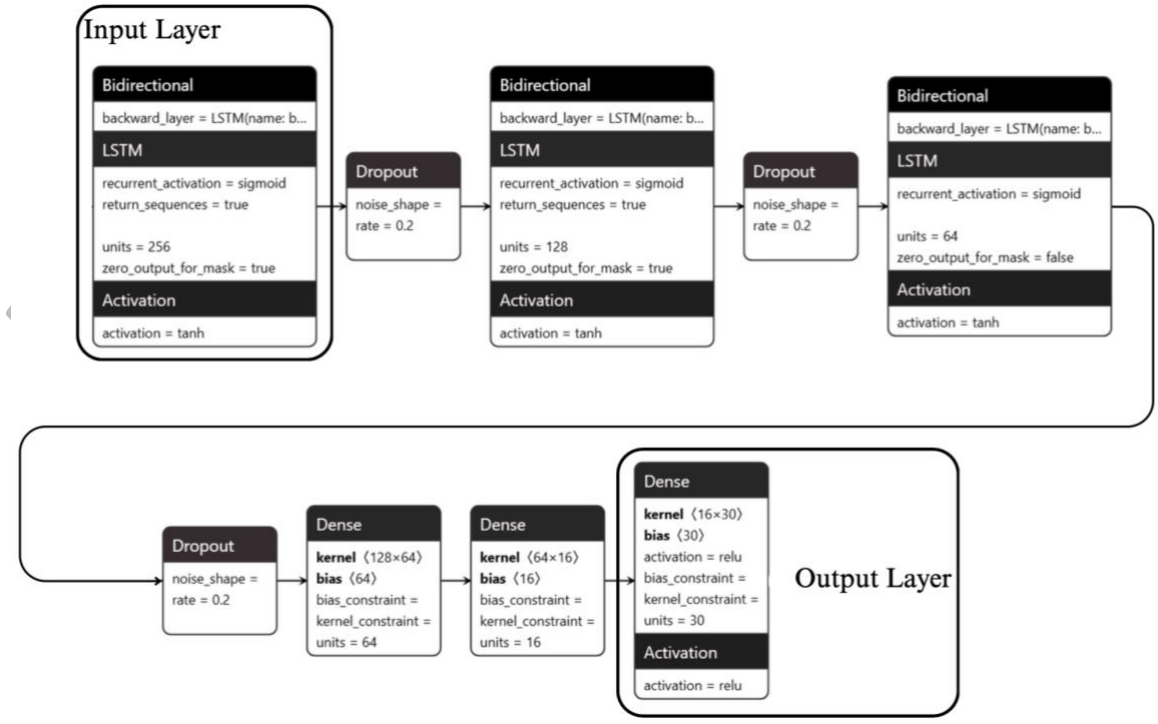


Figure 2. Architecture of the bidirectional LSTM models for regression tasks (Distance and Duration).

outputs via classification probabilities (e.g., Prediction = Regression × Probability) to minimize false positives on non-running days.

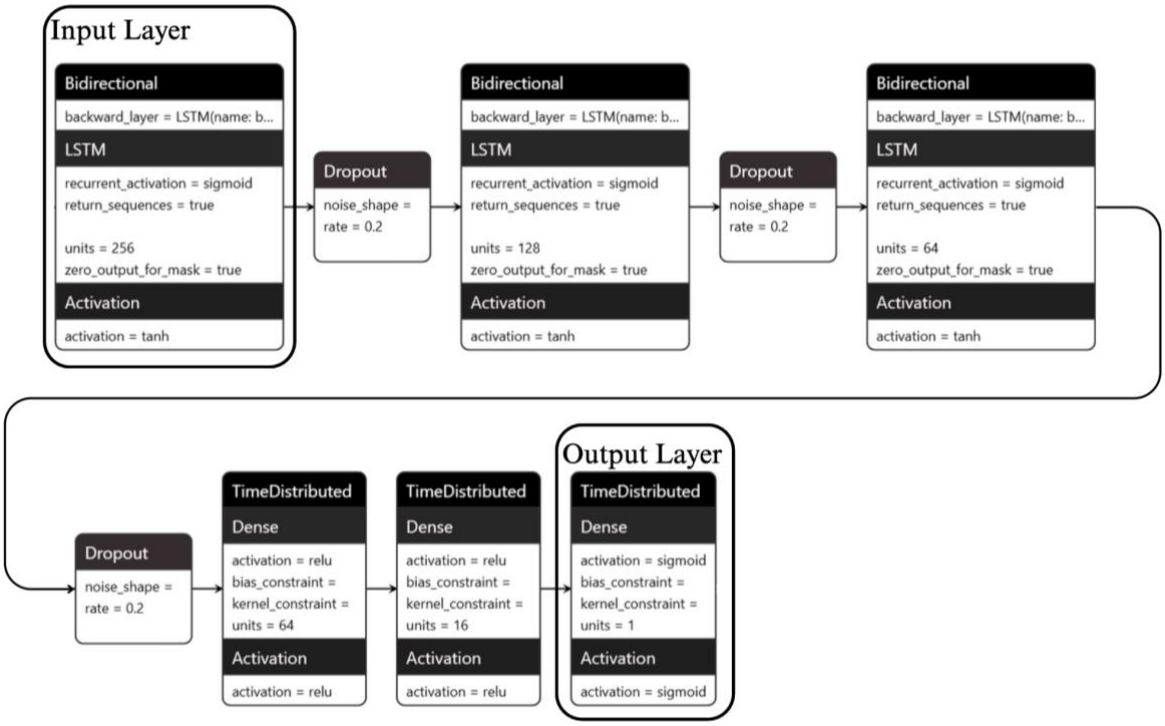


Figure 3. Architecture of the bidirectional LSTM model for binary classification (IsRunning).

3.6 Performance Evaluation

Model effectiveness was assessed using standard performance metrics. Regression models were evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE), while the classification model was assessed using confusion matrix-based metrics: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) (Hossin & Sulaiman, 2015).

3.6.1 Mean Squared Error (MSE)

MSE measures the average squared difference between the actual value (y_i) and the predicted value ($\hat{y}_i$) across n samples. It penalizes larger errors significantly, as defined in Equation (3).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

3.6.2 Mean Absolute Error (MAE)

MAE calculates the average absolute difference between the actual value (y_i) and predicted value ($\hat{y}_i$). In contrast to MSE, MAE assigns equal weight to all errors, providing robustness against outliers, as defined in Equation (4).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \quad (4)$$

3.6.3 Accuracy

Accuracy represents the overall ratio of correctly predicted observations to the total observations, as defined in Equation (5).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

3.6.4 Precision

Precision quantifies the proportion of positive identifications that were correct, indicating the model's reliability in predicting running events, as defined in Equation (6).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

3.6.5 Recall

Recall (or sensitivity) measures the proportion of actual positives correctly identified by the model, as defined in Equation (7).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

3.6.6 F1-Score

JT pre-p

F1-Score is the harmonic mean of Precision and Recall. It is prioritized to ensure a balanced evaluation, especially for imbalanced datasets, as defined in Equation (8).

$$\text{F1} - \text{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

4. Experimental Results

4.1 Deep Learning Model Performance

The predictive capabilities of the system were evaluated using the LSTM architecture described in the methodology. The evaluation is categorized into regression tasks (Distance and Duration) and binary classification (IsRunning).

4.1.1 Regression Models (Distance and Duration)

The regression models were tested on a hold-out dataset of 3,277 athletes. The performance metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE), as shown in Table 2.

Table 2. Performance of Regression Models.

Model	MSE	MAE
Distance Model	42.0293	4.2664
Duration Model	1590.6757	24.4588

The training and validation loss curves for both models are presented, as shown in Figure 4 and Figure 5. The convergence of loss values indicates that the models effectively captured temporal patterns with minimal overfitting.

4.1.2 Binary Classification Model (IsRunning)

The IsRunning model, designed to predict daily activity probability, was evaluated using a dataset of approximately 19.9 million records, as shown in Table 3.

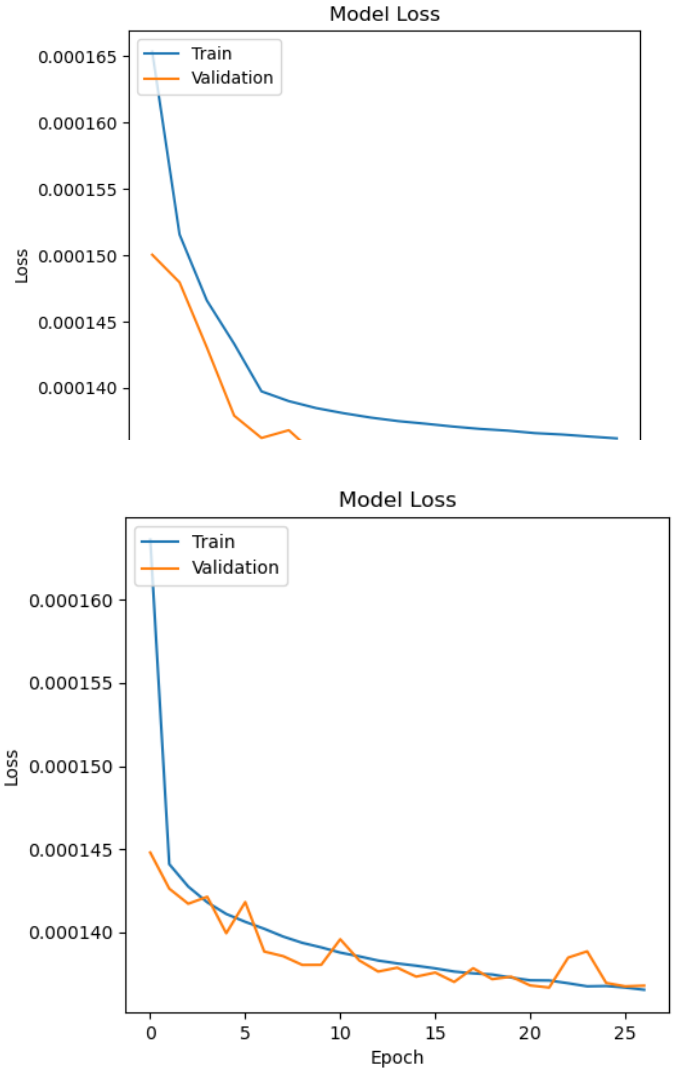


Figure 5. Training and validation loss curves of the Duration Model.

JIT Proof

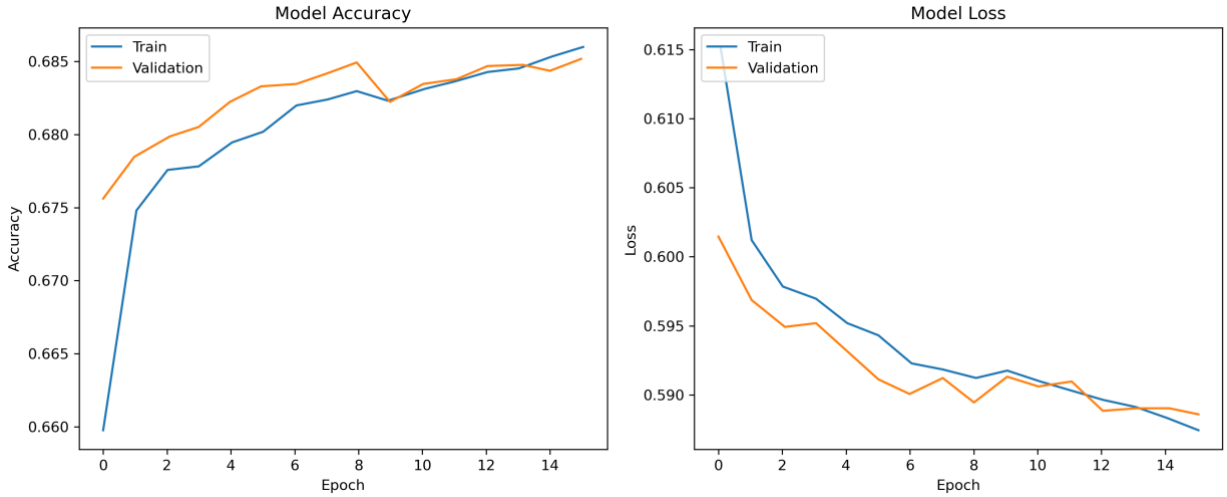


Figure 6. Training and validation metrics (Loss and Accuracy) for the IsRunning classification model.

Table 3. Classification Report for IsRunning Model.

Class	Precision	Recall	F1-Score
0 (Non-Run)	0.68	0.70	0.69
1 (Run)	0.68	0.66	0.67
Accuracy			0.68

To assess model stability during training, the accuracy and loss curves were analyzed, as shown in Figure 6. The model demonstrates steady convergence in both training and validation phases, indicating effective learning of activity patterns without significant overfitting.

The model achieved 68% accuracy with balanced F1-scores (0.69 vs. 0.67), indicating minimal class bias, as shown in Table 3. The confusion matrix, as shown in Figure 7 further validates this, demonstrating moderate consistency in distinguishing between rest and active days.

4.1.3 Integrated Model Performance

To refine the predictions, an Integrated Model was developed by modulating the regression outputs with the classification probability (i.e., Predicted_Value × IsRunning_Probability), as shown in Table 4.

Table 4. Performance of Regression Models.

Model	MSE	MAE
Integrated Distance Model	83.7107	5.3147
Integrated Duration Model	1852.7449	21.1524

As shown in Table 4, the integration yielded mixed results. While the Duration Model's MAE significantly decreased from 24.4588 to 21.1524, the Integrated Distance Model experienced a marginal increase in MAE (from 4.2664 to 5.3147).

4.2 OCR Performance Results

The EasyOCR system was tested using a set of 10 running summary images from various fitness applications to assess the reliability of the automated data collection module. Performance was quantified using the Word Error Rate (WER) metric, as shown in Table 5.

Table 5. OCR Performance and Word Error Rate (WER).

Image Sample	Status	WER (%)
Sample 1 – 2, 4 – 5, 7 - 9	Complete	0%
Sample 3	Incomplete	50%
Sample 6	Incomplete	33%
Sample 10	Incomplete	20%
Average	-	10.3%

Table 5 summarizes the extraction results. The system successfully extracted all key metrics (Distance, Duration, Pace, Date) from 70% of the images (7 out of 10), achieving a perfect WER of 0%. Incomplete extractions were observed in the remaining three samples, with WERs of 50%, 33%, and 20%, respectively, resulting in an overall average WER of 10.3%. These errors were primarily attributed to non-standard fonts and complex background gradients, which interfered with text detection boundaries.

4.3 Usability and Learnability Evaluation

System utility was evaluated through usability and learnability testing conducted with 3 users who were the runners. Figure 8 shows user interfaces, including (A) Login screen, (B) Upload result, and (C) Profile dashboard.

To assess system learnability, the evaluation measured only task duration, ensuring no personal user data was collected. As detailed in Table 6, all tasks achieved a 100% completion rate. The automated OCR process, as shown in Figure 8B, proved highly efficient, averaging 18.02 seconds compared to 31.77 seconds for manual correction. Furthermore, the significant reduction in upload time to 8.53 seconds in the second attempt confirms the learnability objective, demonstrating that users rapidly improved their interaction speed.

JIT pre-proof

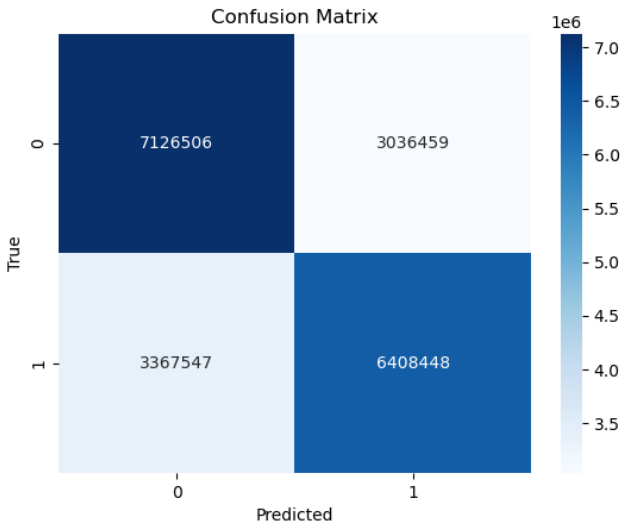


Figure 7. Confusion matrix for the IsRunning binary classification model.

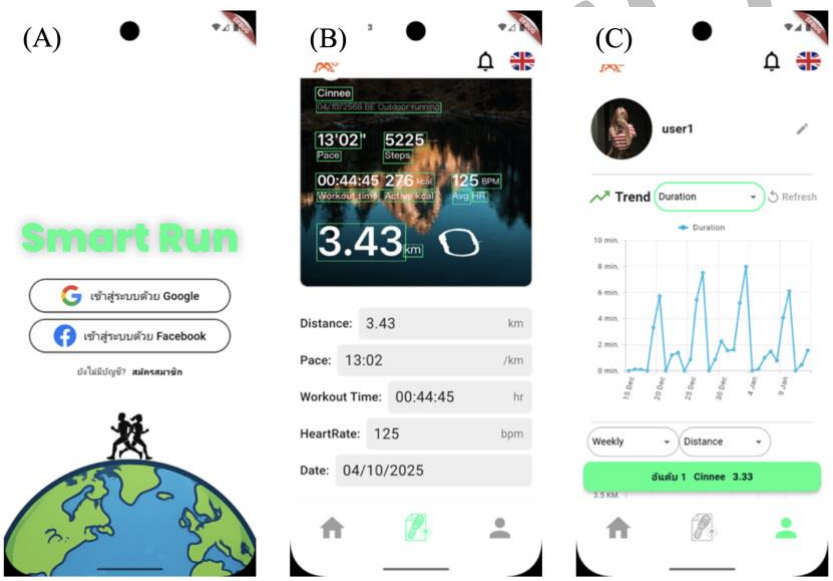


Figure 8. User interfaces of the "Smart Run" application: (A) Login screen, (B) Upload result, and (C) Profile dashboard.

Table 6. Usability and Learnability Evaluation.

No.	Task	Test Result	User 1 (sec)	User 2 (sec)	User 3 (sec)	Average (sec)
	Usability Test					
1	Login via social media (Google / Facebook)	PASS	19.13	25.21	18.21	20.85
2	Complete Personal Profile Data	PASS	22.21	17.15	26.78	22.05
3	Upload Running Stat (Complete Data Flow)	PASS	18.20	13.45	22.41	18.02
4	Upload Running Stat (Edit Incomplete Data)	PASS	33.61	26.24	35.47	31.77
5	Filter Running Events	PASS	8.53	10.44	6.21	8.39
6	View Profile Dashboard (Trend & Leaderboard)	PASS	15.20	32.11	16.20	21.17
	Learnability Test					
7	Re-upload Running Stat (2nd Round)	PASS	8.61	6.24	10.75	8.53
8	Re-visit Profile Stats (2nd Round)	PASS	5.10	22.13	12.11	13.11
9	Edit Personal Data & Check BMI	PASS	21.41	15.78	14.16	17.12

5. Discussion

The results validate "Smart Run" as an effective platform. For the OCR module, while the 10.3% WER indicates limitations with complex backgrounds, the verification interface proved essential. It allowed users to quickly correct errors (~32 seconds), ensuring data accuracy despite algorithmic imperfections.

For forecasting, the model predicted distance better than duration. This is likely because runners often stick to fixed targets (e.g., 5km, 10km), creating consistent patterns, whereas duration varies with daily physical condition. Furthermore, the Integrated Model presented a trade-off: while it improved duration accuracy by filtering non-running days, it slightly increased distance errors due to false negatives. This indicates that while binary classification reduces noise, more detailed user segmentation is needed to further improve precision.

6. Conclusion

This study presented the development of "Smart Run," a mobile application designed to address the challenge of data fragmentation in running analytics. By integrating an EasyOCR-based module with a Bidirectional LSTM network, the system successfully automates data collection and provides personalized performance forecasting. Key contributions include a reliable digitization process with a 70% perfect extraction rate and an effective regression model for distance forecasting (MAE ~4.27 km). Moreover, usability testing confirmed the system's practical utility, demonstrating high learnability and efficient operation times.

Future work will focus on overcoming the current classification limitations by implementing soft clustering techniques. This approach aims to segment users based on distinct running behaviors (e.g., casual joggers vs. marathon runners) to train cluster-specific models, thereby refining prediction accuracy and delivering more relevant personalized insights.

Acknowledgments

This research project was financially supported by scholarship from the Development and Promotion of Science and Technology Talents Project (DPST) under contract number 2-660155-1-SU-66 and 2-680269-1-SU-68. The author also gratefully acknowledge the Department of Computing, Faculty of Science, Silpakorn University for providing the facilities and support for this research.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, June 16-20, 2019, 9365–9374.

Choe, J.P., Lee, S., & Kang, M. (2025). Machine learning modeling for predicting adherence to physical activity guideline. Scientific Reports. 15(1), 5650.

Google. (2024). Flutter - Build apps for any screen. Google. Retrieved 19 May 2025. Retrieved from https://flutter.dev.

Google Developers. (2024). Firebase Documentation. Google. Retrieved 7 July 2025. Retrieved from https://firebase.google.com/docs.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. Neural Computation. 9(8), 1735–1780.

Hossin, M. & Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process. 5(2), 1–11.

Huber, P.J. (1992). Robust Estimation of a Location Parameter. In Breakthroughs in Statistics, S. Kotz and N.L. Johnson, editors. Springer Series in Statistics. Springer, New York, NY, U.S.A.

Internet Engineering Task Force (IETF). (2012). The OAuth 2.0 Authorization Framework (RFC 6749). Internet Engineering Task Force. Retrieved 10 September 2025. Retrieved from https://tools.ietf.org/html/rfc6749.

Jaided AI. (2020). EasyOCR: Ready-to-use OCR with 80+ languages supported. GitHub Repository. Retrieved 21 May 2025. Retrieved from https://github.com/JaidedAI/EasyOCR.

Manikandan, A.V.M., Choudhury, S., & Majumder, S. (2017). Text reader for visually impaired people: Any reader. Proceedings of the IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, September 21-22, 2017, 2389–2393.

Mexwell. (2023). Long-distance running dataset. Kaggle. Retrieved 15 August 2025. Retrieved from https://www.kaggle.com/datasets/mexwell/long-distance-running-dataset.

Morris, A.C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP), Jeju Island, Korea, October 4-8, 2004, 2761–2764.

Pallets. (2024). Flask: User's Guide. Pallets Projects. Retrieved 24 June 2025. Retrieved from https://flask.palletsprojects.com.

Panwar, P., Bhutani, K., Sharma, R., & Saini, R. (2023). A study on calories burnt prediction using machine learning. ITM Web Conferences. 54, 01010.

Schuster, M. & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. 45(11), 2673–2681.

Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39(11), 2298–2304.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research. 15(1), 1929–1958.

Tan Jing Sheng, A., Che Embi, Z., & Hashim, N. (2024). Comparison of Machine Learning Methods for Calories Burn Prediction. Journal of Informatics and Web Engineering. 3(1).

World Health Organization. (2024). Physical activity. World Health Organization. Retrieved 10 December 2025. Retrieved from https://www.who.int/news-room/fact-sheets/detail/physical-activity.

JT pre-proof