FASSI 2023

The Ninth International Conference on Fundamentals and Advances in Software Systems Integration

ISBN: 978-1-68558-096-4

September 25 - 29, 2023

Porto, Portugal

FASSI 2023 Editors

Petre Dini, IARIA, USA/EU

FASSI 2023

Forward

The Ninth International Conference on Fundamentals and Advances in Software Systems Integration (FASSI 2023), held on September 25-29, 2023, continued a series of events started in 2015 and covering research in the field of software system integration.

On the surface the question of how to integrate two software systems appears to be a technical concern, one that involves addressing issues, such as how to exchange data (Hohpe 2012), and which software systems are responsible for which part of a business process. Furthermore, because we can build interfaces between software systems we might therefore believe that the problems of software integration have been solved. But those responsible for the design of a software system face a number of trade-offs. For example the decoupling of software components is one way to reduce assumptions, such as those about where code is executed and when it is executed (Hohpe 2012). However, decoupling introduces other problems because it leads to an increase in the number of connections and introduces issues of availability, responsiveness and synchronicity of changes (Hohpe 2012).

The objective of this conference is to work toward on understanding of these issues, the trade-offs and the problems of software integration and to explore strategies for dealing with them. We are interested to receive paper from researchers working in the field of software system integration.

We take here the opportunity to warmly thank all the members of the FASSI 2023 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to FASSI 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also thank the members of the FASSI 2023 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that FASSI 2023 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of software systems integration. We also hope that Porto provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

FASSI 2023 Chairs

FASSI 2023 Steering Committee
Chris Ireland, OnMove, UK

FASSI 2023 Publicity Chair
Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universitat Politecnica de Valencia, Spain

FASSI 2023

Committee

FASSI 2023 Steering Committee
Christopher Ireland, OnMove, UK

FASSI 2023 Publicity Chair
Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universitat Politecnica de Valencia, Spain

FASSI 2023 Technical Program Committee

Ghaleb M. Abdulla, Lawrence Livermore National Laboratory, USA
Lavanya Addepalli, Universitat Politecnica de Valencia, Spain
Isabela Alves Marques, Universidade Federal de Uberlândia, Minas Gerais, Brazil
Mariia Andriyiivna Nazarkevych, Lviv Polytechnic National University, Ukraine
Vu Nguyen Huynh Anh, Center in Management Information Systems - Université catholique de Louvain, Belgium
Pablo O. Antonino, Fraunhofer IESE, Germany
Imen Ben Mansour, University of Manouba, Tunisia
Dhouha Ben Noureddine, University of Carthage/ University of El Manar, Tunisia
Silvia Bonfanti, University of Bergamo, Italy
Michael Franklin Bosu, Waikato Institute of Technology, New Zealand
Antonio Brogi, University of Pisa, Italy
Evaristo De Jesus Navarro Manotas, Universidad de la Costa, Colombia
Nitish Devadiga, Datarista Inc. / Carnegie Mellon University, USA
Michal Doležel, Prague University of Economics and Business, Czech Republic
Ivanna Dronyuk, Lviv Polytechnic National University, Ukraine
Imane El Alaoui, University of Ibn Tofail, Kénitra, Morocco
Aziz Fellah, Northwest Missouri State University, USA
Ali Mohsen Frihida, University of Tunis El Manar, Tunisia
Youssef Gahi, Ibn Tofail University, Kenitra, Morocco
Eduardo José Gómez Hernández, University of Murcia, Spain
Alan Hayes, University of Bath, UK
Sebastian Herold, Karlstad University, Sweden
Volodymyr Hrytsyk, National University of Lvivska Politechnika, Ukraine
Anca Daniela Ionita, University Politehnica of Bucharest, Romania
Christopher Ireland, OnMove, UK
Yassine Issaoui, University Casablanca, Morocco
Ivan Izonin, Lviv Polytechnic National University, Ukraine
Asharul I. Khan, Sultan Qaboos University, Muscat, Oman
Wiem Khlif, Mir@cl laboratory, Sfax, Tunisia
Elmar Krainz, FH JOANNEUM University of applied Sciences, Austria
Cristiane Lana, Universidade Federal do Espírito Santo (UFES), Brazil
Damian Lyons, Fordham University, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Machine Learning-based Impact Analysis Tool and its Improvement Using Co-occurrence Relationships

Teppei Kawabata
Shibaura Institute of Technology
Tokyo, Japan
ma22045@shibaura-it.ac.jp

Tsuyoshi Nakajima
Shibaura Institute of Technology
Tokyo, Japan
tsnaka@shibaura-it.ac.jp

Shuichi Tokumoto
Information Technology R&D Center, Mitsubishi Electric Corporation
Kanagawa, Japan
Tokumoto.Shuichi@dr.mitsubishielectric.co.jp

Ryota Tsukamoto
Information Technology R&D Center, Mitsubishi Electric Corporation
Kanagawa, Japan
Tsukamoto.Ryota@dy.mitsubishielectric.co.jp

Kazuko Takahashi
Information Technology R&D Center, Mitsubishi Electric Corporation Kanagawa, Japan
Takahashi.Kazuko@dx.mitsubishielectric.co.jp

***Abstract*— In the development of diverted software, impact analysis, which determines the extent of software impact on change requests, is an important task because it greatly affects the quality and efficiency of the development. We proposed a method that machine-learns the modification histories of the projects using word-embedding techniques and multi-label classifiers to accurately generate a ranking list of modification candidates of the software components in order of their sigmoid values. To improve accuracy of the method, this paper proposes to use the multi-label classifier algorism to take co-occurrence between labels into account because of the assumption of the dependencies between the components. Experiments were conducted on actual project data to compare the accuracy of the four algorisms: Convolutional neural networks, BR method, LP method, and RAkEL method. The result shows that RAkEL method, which takes co-occurrence relationships into account and does not over-learn, has the best accuracy among them.**

Keywords—impact analysis; change requests; machine learning; co-occurrence relationships;

I. INTRODUCTION

Software impact analysis is the task of determining the extent to which a change request affects when implemented [1]. Its failure may result in incomplete implementation of change requests or degrades on existing functionalities. Therefore, when many small projects that put small changes on a large source code base concurrently and continuously run, the accuracy and efficiency of impact analysis is crucially important for their development productivity and quality [2].

When conducting an impact analysis to identify the components to be fixed in a source code base for a change request (modification targets), the following two tasks are required: selecting modification candidates and determining modification targets from them. Since the latter task can only be performed by the developers by reviewing the modification candidates, it is important how well the former task can be performed with complete coverage and without waste in order to increase the accuracy and efficiency of the impact analysis.

Requirements traceability is commonly used to seek for modification candidates. It is a discernible association between a requirement and its relating requirements, generally difficult to establish and maintain traceability continuously with a high degree of accuracy. Moreover, it is required to correctly identify the existing requirements affected by the change request before applying the traceability.

Iwasaki et al. [4] proposed and implemented a method for estimating a list of components of a source program as modification candidates directly from a change request by machine-learning the history of changes for the change requests. The implemented tool has two components, one is word embedding part which translates a change request text into a vector form, and the other is machine-learning part which estimates modification candidates from the change request vector. In paper [4], the machine-learning part was implemented using a convolutional neural network (CNN) and evaluated using real project data. The results showed that the method works effectively when there exist many projects each of which implements a small set of change requests for the same source code base.

The tool provides the ranking list of components most likely to be modified (modification candidates) in descendant order of sigmoid value, however it does not provide how to determine the range of modification candidates from which the reviewer determines the modification targets. To determine the range of modification candidates, we set the threshold from the actual data so that it can narrow the range to around 30%. As a result, the rate of missing modification targets for the candidate range was around 23% in case of the above implementation.

In this paper, to improve the accuracy of the tool, we propose to change machine-learning part of the tool from CNN to the other multi- label classification algorithms that take into account label correlations: LP (Label Powerset) and RAkEL (RAndom k-labELsets) [12]. In addition, for comparison of performance, CNN, and BR (Binary Relevance) are used in the experiment. As a result, he RAkEL shows the best results among them.

 ISBN: 978-1-68558-096-4

Section 2 describes the proposed method and its implementation presented in paper [4], Section 3 introduces the algorithms for multi-label classification, and Section 4 describes the experiment to compare the four implementations and show its evaluation results.

II. PROPOSED METHOD AND ITS CNN IMPLEMENTATION

This section describes the experimental data treated in this study and the algorithmic structure of previous studies.

A. Characteristics of the Target Project and its Deliverables

In the software development for diverse and continuously evolving products (multi-product, small-change development [5]), a large number of changes have been made to a source code base to periodically add new features, customizing it for different sets of hardware and various shipping destinations. This type of development often causes many small projects with multiple change requests running in parallel without sufficient human resource with sufficient knowledge on the source code infrastructure to perform impact analysis accurately and efficiently.

The target project group adopts a derivative development method called XDDP (eXtreme Derivative Development Process) [6], in which one change design document is supposed to be created for each change request.

This document describes the following items.

- Change Request ID
- Requirements (natural language)
- Mounting
- Details of changes to software method design specifications
- Modification details regarding the module of the software detailed design specification
- Description of changes made to the source code, including names of components and modules that have been modified

About 30 projects occur every year, and about 10 change requests are made per project in average.

The input change request text is written in Japanese, having 20 to 400 characters, and the output is 32 components, which the source code based has.

B. Proposed Method

The proposed method learns a large number of change design documents to estimate modification candidates directly from new change requests [4].

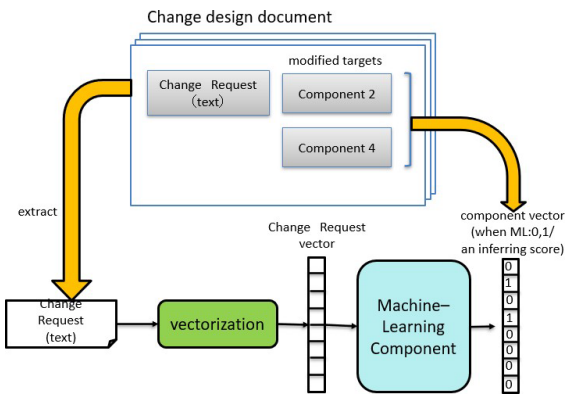

Figure 1 Configuration of the implemented tool.

As shown in Figure 1, for each change request document, a change request text is extracted, and then a vector of change request text is created using a word-embedding technology. On the other hand, a component vector is created from the information on the modified modules in the source code base corresponding to the change request. Each index of the vector is uniquely corresponding to some component in the source code base, whose value is either 0 or 1 (1 means modified and 0 means unmodified).

At the time of estimation, the proposed method outputs the ranking list of components most likely to be modified for a new change request text.

Compared to the impact analysis using traceability, the proposed method has the advantage of being able to select modification candidates directly from a new change request without burdening development activities.

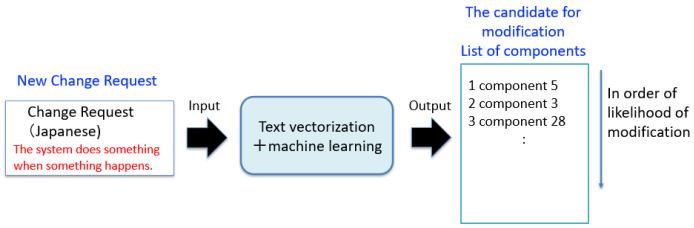

Figure 2 Proposed Method for Estimating Modified Candidates.

C. Algorithm Structure of Previous Studies

1) Text vectorization

When vectorizing a text, the text is decomposed into words by applying the morphological analyzer mecab [7]. The decomposed words are processed in three stages: extraction of words to be used from all the words (word extraction), vectorization of the words, and vector integration. The resulting vector has 100 dimensions.

In a previous study, we tried three implementation methods shown in Table 1 and as a result found out that noun selection + doc2vec [8] (Implementation 3) produced the best results.

TABLE I: IMPLEMENTATIONS EVALUATED IN PREVIOUS STUDY

Implementation No	Word extraction	Word vectorization	Vector integration
I1	Noun selection	word2vec (skip-gram)	Simple averaging
I2	Full selection	doc2vec	
I3	Noun selection	doc2vec	

2) Machine learning

The machine-learning part can be seen as the multi-label classifier since it determines whether the 32 vectors of values are 0s or 1s for a vector of change request text. To implement the tool, convolutional neural networks (CNNs) have been used as a multi-label classifier.

Figure 3 shows the structure of the implemented CNN. The input is a 100-dimensional vector of change request text, and the output is a 32-dimensional vector of component lists. The reason the number of components output is 32 is that the number of components in the data used in the experiment is fixed at 32.

 ISBN: 978-1-68558-096-4

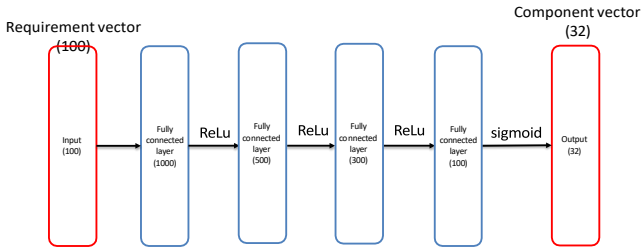

Figure 3 Adopted structure of CNN.

The parameters of the CNN are as follows.

- Intermediate layer: 4 (1000, 500, 300, 100)
- Number of epochs: 50
- Batch size: 50
- Learning rate: 0.1
- Function on output: sigmoid

D. Reassessment of Previous Studies and Issues

In the previous study [4], a ranking list of modification candidates is ordered by the likelihood to be modified to a change request. The output list contains a mixture of modification targets and the others. The modification targets are the components that need to be modified by the change request.

The previous study used two metrics to evaluate its performance: coverage range ratio and accuracy in the coverage range, where the coverage range is up to the position where the last modification target appears in the list.

However, prior research has not provided a method for determining which components need to be reviewed. To do this, we devise a method that determines a threshold on the sigmoid value to determine the range to be reviewed, where the threshold is to be determined from the actual data to be a specified range ratio. In addition, we defined three metrics shown in Figure 4 to evaluate the performance of the method.

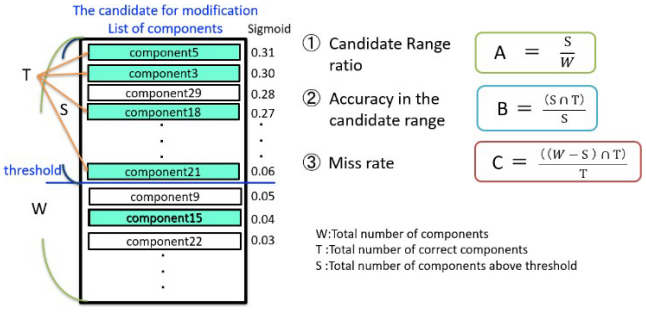

Figure 4 Measurements to evaluate effect on impact Analysis.

1. Candidate range ratio: percentage of components with higher sigmoid value than the threshold out of the number of all components.
2. Accuracy in the candidate range: percentage of modification targets out of components in the candidate range.
3. Missing rate: percentage of the modification targets beyond the candidate range out of all the modification targets.

Table 2 shows the metrics measured for the data of the previous study.

TABLE II THREE METRICS FOR THE PREVIOUS STUDY

Threshold	A	B	C
0.06	30.0%	18.0%	23.0%

In the Table 2, the threshold is 0.06 when A is set to 30%. In case, the tool of the previous study resulted in B of 35% and C of 23%. The problem is that C is considerably high. A higher missing rate may cause bugs in the program because it increases the likelihood of leakage of reviewing. Therefore, it is necessary to reduce the missing rate for practical use.

E. Improvement Targets

The goal of this study is that the candidate range ratio is less than 30% and the missing rate is 5% or less. The reason we set the goal is that we want to keep the missing rate within the 2σ interval from the viewpoint of quality assurance, obtaining a certain level of effort reduction of reviewing tasks.

III. ALGORITHMS FOR MULTI-LABEL CLASSIFICATION AND CO-OCCURRENCE RELATIONSHIPS

This section explains the reasons for focusing on co-occurrence relationships and the methods that take co-occurrence relationships into account.

A. Co-occurrence relationships in the source code base

To improve the missing rate in the previous study, we focus on the architectural dependencies between the components in the source code base. Such architectural dependencies include:

- Call Relationships
- Resource sharing relationships (communication, memory, I/O)
- File read/write relationships
- Inheritance relationships
- Include relationships

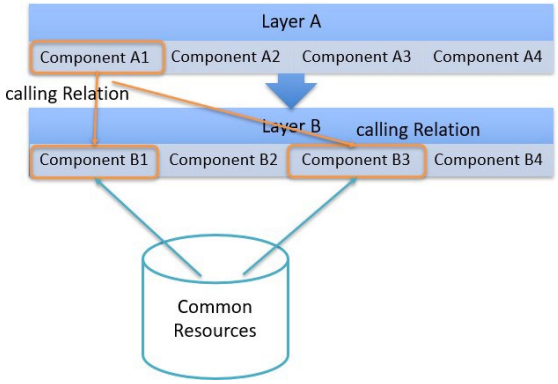

Figure 5 Dependencies between components.

In particular, the target source code base has a layer structure as illustrated in Figure 5, having:

- Calling relationships may occur between the components of adjacent layers.

- Components in a specific layer (Layer B in Figure 5) handle a common resource, causing indirect dependencies.

Concerning this observation, we hypothesize that components having dependencies are often modified together. If it is true, we can improve the machine-learning part by applying multi-label classification algorithms that take into account label correlations to it, which may increase its performance.

B. Algorithms for handling multi-label classification

As mentioned earlier, machine-learning in this study is attributed to the problem of multi-label classification, in which multiple labels are assigned to a single object [7].

The major difference between multi-label classification and single-label classification is that it is expected to improve accuracy by using the co-occurrence relationships between labels in the prediction process.

In this study, in order to incorporate co-occurrence relationships among outputs, several multi-label classifiers are examined, with a combination with the Support Vector Machine (SVM). SVM is a supervised learning algorithm [8] that can be used for classification and regression problems such as natural language processing and speech recognition.

The Binary Relevance (BR) method is one of the representative methods for multi-label classification (without considering correlations between labels), predicting labels by transforming a multi-label classification into multiple single-label classification. In detail, the BR method creates a binary classifier for each label and outputs the sum of the classifier results [9]. For our problem, each element of the component vector is trained with the input sentences vector (Figure 6).

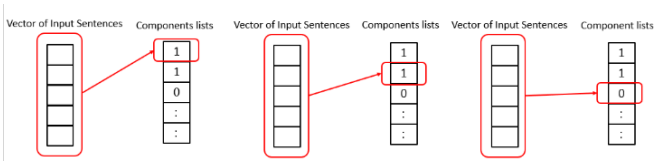

Figure 6 Configuration of the learning algorithm for the BR method.

C. Algorithm to model co-occurrence relationships

1) Label Powersets method (LP method)

Label Powersets method (LP) is one of the basic algorisms for multi-label classification considering the correlation between labels.

The LP method treats each element of the power set of labels as a class, transforming multi-label classification into multi-class classification. A power set is all possible

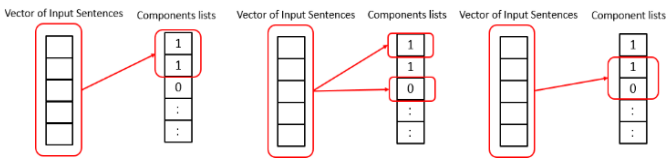

Figure 7 Configuration of the learning algorithm for the LP method.

Combinations, for example, a power set of labels 1, 2, and 3 are ϕ, {1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3}, {1, 2, 3}. The LP method use all the set except ϕ as classes, classifying an input into one class.Figure 7 illustrates what pattens are to be learned.

The LP method calculates the probability of occurrence of a label from the sum of the probability of occurrence of the classes in which the label appears, shown in Figure 8.

Model	Probability of appearance of class	Estimation Results by Label				
		label1	label2	・・・	label31	label32
set 1 classifier	0.1	1	0	・・・	-	-
set 2 classifier	0.2	-	-	・・・	1	0
:	:	:	:	:	:	:
set N classifier	0.3	1	0	・・・	0	1
result	result	0.1*1 + 0.3*1	0	・・・	0.2*1	0.3*1

Figure 8 Estimation results for each label.

Although the LP method has the advantage of prediction based on co-occurrence relationships among labels, it has some disadvantages:

- The computational complexity increases exponentially with the number of labels.
- The number of classes increases, resulting in overlearning when the number of data is small.

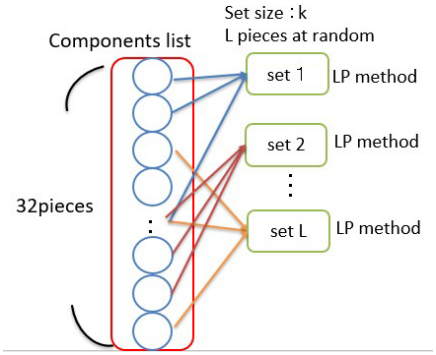

Figure 9 Creating a subset from a set of labels.

2) Random k Labelsets method (RAkEL method)

As mentioned above, the LP method has some disadvantages when the number of labels increases. The RAkEL method [10] was proposed to conquer them.

Model	Estimation Results by Label				
	label1	label2	・・・	label31	label32
set 1 classifier	1	-	・・・	0	-
set 2 classifier	-	0	・・・	1	-
:	:	:	:	:	:
set L classifier	0	0	・・・	-	1
result	T_1/M_1	T_2/M_2	・・・	T_{31}/M_{31}	T_{32}/M_{32}

T1: Number of cells whose estimated result is 1 , Mi: Number of cells with estimated resul

Figure 10 Estimation results by label.

 ISBN: 978-1-68558-096-4

The RAkEL method first randomly creates label subsets of size *k* for the input label set, and then applies the LP method to each subset, as illustrated in Figure 9.

The classification results for each subset are then integrated to predict the label, as shown in Figure 10.

The RAkEL method is an algorithm with high potential for improving accuracy compared to the LP method because it can significantly reduce the computational complexity of LP calculations for subsets compared to the LP method for the whole set, and it can also reduce the bias in the distribution of each class value.

IV. EVALUATION AND EXPERIMENT

This section describes the experimental results and evaluation of the proposed method.

A. *Purpose of the experiment*

We consider that the applicability and accuracy of the algorithms described in previous section will differ depending on the nature of the problem domain and the number of available training data. To examine them, we apply the BR, LP, and RAkEL methods to the machine learning part of the proposed method and conduct an experiment to compare their accuracy using the same project data.

Our research question is whether the algorithms that take co-occurrence relationships into account may improve the accuracy for this problem or not. To answer this question, we select the following four implementation of the machine-learning part for comparison:

1. CNN (implementation in [4], already shown in Table 2)
2. BR with SVM (no consideration on co-occurrence)
3. LP with SVM
4. RAkEL with SVM

This will allow us to evaluate whether algorisms considering co-occurrence relationships improve accuracy, investigating effects of the LP method's disadvantages. Furthermore, by presenting the measurement results of the CNN-based classifier, we will evaluate tow what extent they improve the accuracy from the previous study.

B. *Experimental data*

This experiment uses data from 405 change design documents provided. The data was divided into training and test data at a ratio of 4:1, with 324 books used as training data and 81 books used as test data (Figure 11). This sequence of experiments were conducted five times with other combinations, and the average of the results was calculated as the experimental result.

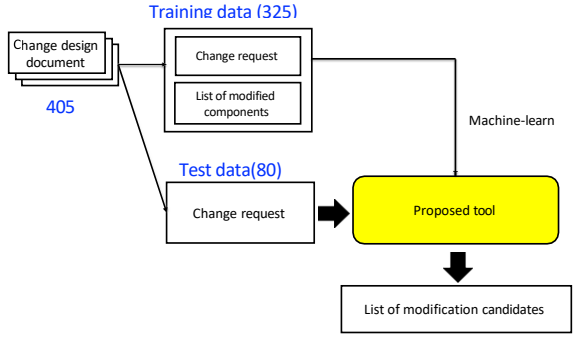

Figure 11 Data size used for the experiment.

C. *Experimental Methods*

The experiment was conducted according to the following procedure.

1. Change request sentences are vectorized by noun extraction + doc2vec and features are extracted.
2. The machine learning component is configured using Scikit-learn. BR/LP/RAkEL methods as multi-label classifiers (k=3) ＋Implemented three SVMs: + SVM.
3. For the five sets of experimental data shown in Figure 11, the following were performed and averaged to calculate the accuracy:
   a) Create a training model from training data.
   b) Estimation of labels is performed on the remaining test data.

TABLE III: BR METHOD + SVM Values

Threshold	Percentage of candidate Range	Accuracy in the candidate range	Missing rate
0.04	38.1%	14.7%	12.3%
0.05	33.5%	17.1%	15.2%
0.06	*29.4%*	*19.1%*	*17.1%*
0.07	25.7%	21.0%	20.5%
0.08	23.1%	22.4%	23.7%
0.09	21.1%	23.1%	28.1%
0.1	19.4%	24.5%	29.9%

TABLE IV: LP METHOD + SVM

Threshold	Percentage of candidate Range	Accuracy in the candidate range	Missing rate
0.04	46.0%	16.8%	10.0%
0.05	39.9%	18.7%	13.4%
0.06	35.8%	19.9%	16.7%
0.07	32.8%	21.0%	19.7%
0.08	*29.7%*	*22.2%*	*23.0%*
0.09	26.6%	23.7%	26.4%
0.1	24.1%	25.0%	29.9%

TABLE V: RAkEL method ＋SVM

Threshold	Percentage of candidate Range	Accuracy in the candidate range	Missing rate
0.04	41.4%	18.7%	9.6%
0.05	36.5%	20.8%	11.3%
0.06	32.6%	22.7%	13.5%
0.07	*29.5%*	*24.5%*	*15.6%*
0.08	27.2%	26.2%	16.7%
0.09	25.2%	27.6%	18.8%
0.1	23.1%	29.3%	20.9%

 ISBN: 978-1-68558-096-4

D. Evaluation methods

Three measures were obtained: the candidate range ratio, which indicates effectiveness of narrowing the review range; the accuracy in the candidate range, which indicates amount of waste in the review process; and the missing rate, which indicates adequacy of determining candidate range.

The sigmoid threshold was moved in 0.01 increments until the candidate range ratio over 30 percent. The threshold value where it is the closest to 30 for each algorism is selected for comparison.

E. Experimental results

Table 3-5 show the values of three measures respectively for method 2-4. The values are shown in the range of 0.04 to 0.1 for sigmoid values, in line with previous studies. The value when it is closest to 30% is shown.

Table 6 compares the accuracy of four methods (including CNN's in Table 4) around a candidate range ratio of 30%.

TABLE VI: COMPARISON RESULTS OF ALL METHODS

method	Candidate range ratio (threshold)	Accuracy in the candidate range	Missing rate
CNN(Previous research)	30.00% (0.06)	18.00%	23.00%
BR＋SVM	29.10% (0.06)	19.10%	17.10%
LP+SVM	29.70% (0.08)	22.20%	23.00%
RAkEL+SVM	29.50% (0.07)	24.50%	15.60%

The results of our analysis are:

- The accuracy of the error rate was improved by 5.9% when comparing the BR method + SVM with the conventional method (CNN). This result indicates that SVM is more accurate than CNN for this problem.
- When comparing the BR and LP methods, the LP method was less accurate than the BR method, which is not the expected result because it must have superiority of the method that takes co-occurrence relationships into account. This is most likely due to overlearning, as the number of output labels is as large as 32, resulting in a huge number of combinations.
- The RAkEL + SVM method is the most accurate of the above methods, improving the missing rate by 1.5 points and the accuracy in the candidate range by 5.4 points compared to the BR method (improving 7.4 and 6.5 to CNN respectively). This result indicates that the RAkEL method did not cause overlearning problems, showing that the co-occurrence relationship is effective in improving accuracy to some extent.
- The highest accuracy of the RAkEL method was 15.6%, and the target accuracy of 5.0% or less could not be achieved.

V. CONCLUSION

In this paper, we compared and evaluated a total of four algorithms: two that take co-occurrence relationships into account (LP and RAkEL) and two that do not (CNN and BR). The results of the experiment show that the RAkEL method considerably improves the accuracy from the CNN in the previous study.

Future work includes further improvement of algorisms such as the application of the improved RAkEL method (overlapping version). Further validation of the effectiveness of the proposed method by applying it to another dataset is also needed, including exploring the necessary size of historical data to obtain a certain accuracy.

REFERENCES

[1] S. Sikka, A. Dhamija, Software Change Impact Analysis, BookRix, 2020.

[2] Bohner, Impact analysis in the software change process, a year 2000 perspective, Proceedings of International Conference on Software Maintenance, 1996, pp. 42-51.

[3] ISO/IEC/IEEE 24765, 2017 Systems and software engineering - Vocabulary.

[4] H. Iwasaki, et al, A Software Impact Analysis Tool based on Change History Learning and its Evaluation, ICSE-SEIP '22, May 21 – 29, 2022, Pittsburgh, PA, USA.

[5] N. Motoi, T. Nakajima, and N. Kuno, A case study of applying software product line engineering to the air conditioner domain, Proceedings of the 20th International Systems and Software Product Line Conference, 2016, pp.220-226.

[6] K. Kobata, E. Nakai, and T. Tsuda, Process Improvement using XDDP - Application of XDDP to the Car Navigation System, 5th World Congress for Software Quality, Shanghai, China, November 2011.

[7] T. Kudo, K. Yamamoto, and Y. Matsumoto, 2004 Applying conditional random fields to Japanese morphological analysis, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, volume 2004.

[8] Q. Le, T. Milkolov, Distributed representations of sentences and documents, International conference on machine learning, pp.1186-1196, 2014.

[9] Y, Kosuke, et al, "Interdependence Model for Multi-label Classification." International Conference on Artificial Neural Network*s*. Springer, Cham, 2019.

[10] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 1999.

[11] G. Tsoumakas, and I. Katakis, Multilabel classification, An overview, Int J Data Warehousing and Mining, Vol. 2007, pp.1–13.

[12] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multi-label classification," IEEE Transactions on Knowledge and Data Engineering, vol. 99, no. 1, 201.

 ISBN: 978-1-68558-096-4

Evaluating Usability of Artificial Intelligence (AI) Based M-Health Applications Through Cognitive Walkthrough

Majed A. Alshamari
Department of Information Systems, College of Computer Sciences and Information Technology
King Faisal University
Hofuf- Kingdom of Saudi Arabia
Email: smajed@kfu.edu.sa

***Abstract*—Artificial Intelligence (AI) technology has been adopted and employed in healthcare section to develop applications for providing various healthcare services. However, the effectiveness of these apps depends on their usability, which is a critical factor in their success. One approach to evaluating the usability of these apps is through a cognitive walkthrough. In our study, we aimed to evaluate the usability of AI-based features in 3 mHealth applications, including Ada, Babylon, and Ornament. We conducted a cognitive walkthrough by providing a list of tasks in order to carry out the process. After each task completion, evaluators were presented with questionnaire to assess the application's usability attributes. A total of 27 distinct problems were identified. The highest number of problems were related to health information and symptom checking features. The reported severity of identified issues in Ada, Babylon and Ornament are 7.4, 8.0 and 4.2 respectively. Some of the identified usability problems are irrelevant health information, limited disease enlistment, no search option, tiresome navigation, unsatisfactory results, and delayed responses. These issues impact effectiveness, and efficiency of AI models, and ultimately user satisfaction, thus, highlighting the need to improve AI based mHealth applications' functionality and design. Further, the evaluators provide recommendations on these identified problems.**

Keywords- AI based mobile applications Introduction; Artificial Intelligence; Cognitive Walkthrough; mHealth application; Usability Evaluation.

I. INTRODUCTION

Nowadays, everyone irrespective of their age can access smart devices including smart televisions, tablets, phones, and other internet-connected devices because of digital media. Every day, thousands of apps, with a wide range of functions, are added to dedicated (iOS and Android) app stores (Apple App Store and Google Play Store), and this number is constantly growing [1]. In the past decade, the health industry has seen phenomenal growth and pushed healthcare delivery to new levels. Therefore, m-health is becoming an essential sector for delivering and spreading health in our society as a whole [2]. The mobile health (mHealth) app market is anticipated to develop at a compound annual growth rate of 17.7% throughout the forecast period, according to the most recent report by Grand View Research, Inc. [3], reaching US $149.3 billion by 2028. Users' interest in mHealth applications has grown significantly over the past decades, making healthcare a significant category in these mobile app catalogs. According to research, up to 34% of smartphone owners have at least one health app loaded on their mobile devices [4]. Also, the usage of artificial intelligence (AI) in mobile apps for healthcare systems, finance, and entertainment has increased primarily due to smartphones and tablets [5]. In this era of rapid technological development, people from all walks of life utilize artificially intelligent mobile applications (apps) on a global scale. Conclusively, AI is progressively playing a larger role in people's daily lives [6] [7].

Using AI-based applications in healthcare is of particular importance to patients; therefore, it is important that their use does not harm them, but rather benefits them. Thus, AI systems should provide patient satisfaction across multiple healthcare environments and be effective and efficient [8]. Hence, by examining the usability of mobile health apps, we can uncover issues, help redesign systems, spend less time and money, and improve user acceptance [9]. Effectiveness, efficiency, learnability, ease of use, and user satisfaction considered some of the most common usability attributes when defining the usability of system. An often-used analytical method of usability evaluation is Cognitive Walkthrough (CW) [10].

Lewis and Wharton developed cognitive walkthrough for evaluating the usability of interfaces using theory-based evaluation [11]. It is employed to identify problems and generate proposals about their causes. Learning through CWs is aimed at simplifying learning, especially through exploratory learning. In medical equipment evaluation, CW is used to evaluate depression screening models [12], Nurse information systems [13], Diabetes management systems [14] and other healthcare systems. It is advantageous to use CW in healthcare since it can be used to identify important usability problems quite easily, quickly, and cheaply when real usability testing is not feasible [15]. This paper aims to contribute in the identification of critical issues in mobile health applications that affect the usability attributes and that impact the adoption and effective use of AI applications in healthcare.

The rest of the paper is organized as follows: In Section 2 we discuss the previous publications which evaluates the AI based system. Section 3 is the detailed methodology of our evaluation process. In Section 4, we present the qualitative

 ISBN: 978-1-68558-096-4

and quantitative findings from evaluation, followed by Section 5, which is the conclusion of our research work.

II. LITERATURE REVIEW

There are nearly 165,000 mobile health (mHealth) apps available in the Apple iTunes and Android app stores in the United States [16], which are used by two-thirds (66%) of Americans [17] Survey Finds 66% of Americans eager to leverage digital tools to manage personal health. Many mHealth apps have designed with little input from end users, and they continue to expand despite limited evidence of user engagement [18][16] Applications are routinely created with low-quality designs and with insufficient consideration of end-user demands. Such applications may be challenging to use, misunderstood, or underutilized, and may ultimately fall short of their objectives [19] [20] Apps must therefore guarantee quality and offer the necessary functionality. This emphasizes how crucial it is to assess usability of mHealth applications. Also, medical technology focuses more on usability than user experience [21]. This section provides a summary several research studies that have been performed to calculate usability of mHealth and mobile systems through the usability evaluation methods.

A mobile accounting software has been developed by [22] using Rapid Application Participatory Development (RAPD) method. Further, they evaluated the usability of accounting software using a cognitive walkthrough performed by 16 participants. Their objective was to identify the effect of COVID-19 on the usability of new software. Their research identifies new factors that influence application usability, including user experience, remote work, security, privacy, internet speed and Artificial Intelligence. [23] proposed a conceptual model names " GenDAI", which is an AI assisted laboratory Diagnostic Solution for the genomic applications. In GenDAI, the AI-driven AI2VIS4BigData abstract architecture for metagenomics is combined with the CRISP4BigData-based model for the gene expression diagnostics. An evaluation of this conceptual model was conducted by partnering with small medical laboratory of ImmBioMed GmbH & Co. KG in Heidelberg, Germany. Platflow was developed to perform analysis on the raw data and was evaluated through cognitive walkthroughs. Preliminary study results indicate that there are several areas in laboratory workflow that could be automated.

A depression-screening model has been evaluated by N. Fasihah Jamaludin [12] to examine how effective it was at addressing adolescent motivation during gamification-based depression screening, through a cognitive walkthrough. The evaluation was conducted by five respondents with expertise in adolescent counseling and human-computer interaction. According to the analysis, all respondents gave positive feedback on the sets of tasks provided. These results confirmed the model's usability in detecting depression through the cognitive walkthrough. They concluded that the model might be used as a blueprint for creating a real depression screening system. A. S. Dahri [24] investigates how well the mobile health application "mHealth" is used by patients by accessing their satisfaction with their tasks. 15 patients completed tasks on task success rate, mistakes, efficiency (time spent), and satisfaction using System Usability Scale (SUS) and the International Organization for Standardization (ISO) 9241-11 standard criteria. Effectiveness was measured in terms of how many users have successfully completed task, while efficiency was measured in terms of time taken by each task to get completed. The findings of this study showed that finding a medical professional was the most challenging step for users and registering was the easiest task. The usability scores in this study are also influenced by educational level and mobile expertise.

In addition, M. N. Islam et al [25] developed a mobile-based solution "Muktomon" which means open one's mind, for providing mental health support to the people of Bangladesh. This application provides virtual therapy through videos and audio, a chatbot service for mental health assistance. They evaluated the usability by conducting a system usability survey and pots questionnaire from 37 participants. Their application got SUS score of 79.875%, which means acceptable system in terms of usability. In the context of the COVID-19 pandemic, the application proved useful and usable for improving mental health. N. A. Zaini et al [26] designed a low-fidelity APi prototype of a game to provide fire safety education to children. They used interactive learning as a key to promoting preschool children's knowledge of fire safety basics. APi prototypes were designed based on the user requirements of preschool children focused on cognitive, psychomotor, and behavioral aspects. A small group of 6 people including professional designers and developers. They conducted the cognitive walkthrough evaluation to evaluate the usability and learnability of the APi interface. Participants evaluated prototype on color, background theme, font, and consistency in design etc. From the findings of the cognitive walkthrough, they designed the high-fidelity prototype of APi interface for fire safety education.

The cognitive evaluation method has been used by [13] to evaluate the usability of a user interface of a Nursing Information System (NIS). The system was evaluated by five evaluators according to given scenarios and the problems identified were assigned to usability attributes. Evaluators also determined the severity of each identified problem. M. Georgsson [14] proposed a technique called user-centered cognitive walkthrough, to address the flaws of the original cognitive walkthrough. They also perform a preliminary validation using the think-aloud protocol to gauge the method's efficacy, and user acceptability in a study with diabetes patient which are users of a mHealth self-management application. They divided the Diabetes patients into 2 groups, one as UC-CW and the other as think-aloud (TA) groups at the University of Utah Health in the United States. They identified 26 different usability problems (heuristics violation) with UC-CW and 20 usability problems using the think-aloud method, in Recall and Recognition, Consistency and Standards, and Match between System and Real world. The study reported that UC-CW is an effective method for finding usability problems than TA becuase patients' diseases required customized qualities that could not

 ISBN: 978-1-68558-096-4

be determined by TA. [27] proposed a study which compares two expert-based evaluation methods (Heuristic Evaluation & Cognitive Walkthrough) in a nursing module of a Hospital Information System (HIS). Five evaluators use the system and identifies 104 problems with the heuristic method and 24 usability problems with cognitive walkthrough method. They reported a significant change between severity of recognized usability problems and the number by these methods. As a result of the cognitive walkthrough, issues of learnability, efficiency, and memorability have been identified, whereas as a result of heuristic evaluation, issues of effectiveness, satisfaction, and errors have been identified. methods.

A usability test involving 18 healthcare professionals has been conducted by [28] to evaluate the effectiveness of an electronic health record (EHR) display prototype for emergency medicine. Participants were asked to complete 2 questionnaires for rating usability, usefulness, and effectiveness. Study findings emphasize the need for user-centered design when developing EHR systems for emergency medicine. [2] developed and evaluated an e-health prototype with five health professionals including information system experts and six health consumers. The Post-Study System Usability Scale (PSSUS) was modified and adapted by the authors, who developed the post-Study e-Health Usability Questionnaire (PSHUQ), which consists of 19 items describing five characteristics of system usability: easy learning, functional adequacy, rapid acquisition of usability experts and several different user groups, rapid completion of work, and high-quality online documentation. A number of users have provided feedback on the system, suggesting improvements and recommendations for future enhancements. The most common suggestion was that consumers' personal information should be kept confidential and secure. Moreover, optimization of resource utilization and quality are desired, along with meeting consumer demands.

TABLE I. LITERATURE REVIEW

Reference #	Objective	No. of Participants	Evaluation method	Evaluated App
[22]	Developed accounting software using RAPD Identified new usability factors	16	Cognitive walkthrough, interviews	Accounting mobile app
[23]	Genome diagnostic tool for laboratory use Developed an application for analyzing results	-	Cognitive walkthrough, on-site visits, interviews	Platflow Tool
[12]	Usability evaluation of the depression screening model	5	Cognitive walkthrough	Gamification Model
[2]	Developed and evaluated an e-health prototype	11	Post-Study System Usability Scale	Heal-me.co
[25]	Development and Usability evaluation of Mental Health care app	37	SUS, Interviews	Muktomon
[13]	Usability evaluation of an information system	5	Cognitive Walkthrough	Nursing Information System
[27]	A comparative study to evaluate usability and learnability of a system with different methods	5	Cognitive Walkthrough +Heuristic Evaluation	Health Information System
[24]	Investigates UE of the Mobile Health application by patients' task performance evaluation and satisfaction	15	SUS, ISO 9241-11	mHealth
[26]	Develop and evaluate the	6	Cognitive Walkthrough	Fire Safety Education

	usability of the prototype for preschool children			
[14]	A case study to evaluate the usability of a mobile-based healthcare system	12	Cognitive Walkthrough	Diabetes self-management application
[28]	Evaluate the usability of Emergency Medicine HER Prototype	18	Questionnaires	Electronic Health Record Display
Proposed Work	Evaluate usability and learnability of AI applications in healthcare	15	Cognitive Walkthrough	Ornament, Ada Health, Babylon Health

The proposed work evaluates the effectiveness, efficiency, ease of use and satisfaction of 3 AI applications (Ada, Babylon and Ornament) in healthcare using cognitive walkthroughs. A significant contribution of this study will be the identification of critical issues in these applications that affect the usability attributes and that impact the adoption and effective use of AI applications in healthcare, and they will contribute to knowledge of usability and learnability. In order to develop more user-friendly AI applications that are easy to use, learn, and adopt in healthcare, the study aims to provide useful recommendations from experts.

III. MATERIALS AND METHODS

In this section, we present the methodology that was used to evaluate the usability of AI-based features in three mobile health applications: Ada Health, Ornament, and Babylon Health. To assess the effectiveness, efficiency, satisfaction, and ease of use of these applications, we used the cognitive walkthrough evaluation method. We chose these three applications because they are well-established and widely used in the healthcare industry and each provides unique AI-based features. Further, we recruited a team to conduct the evaluation. We then conducted a survey that included both qualitative data, that are problems and suggestions, and yes/no responses. The complete workflow diagram is shown in Figure 1.

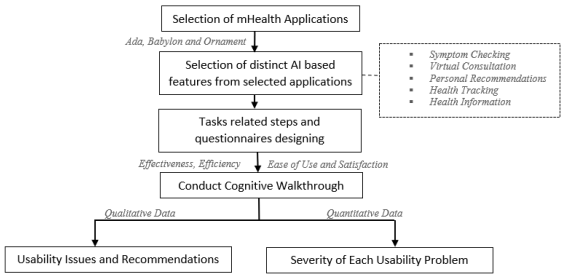

Figure 1. Complete Workflow Diagram of conducting CW to Evaluate usability 3 Applications.

A. *Cognitive Walkthrough Evaluation Method*

A group-based expert approach called CW was created by Polson and Lewis and is based on theories of the cognitive exploratory learning or users' capacities to understand through their activities [29], [30]. Experts identify system flaws using CW by simulating users' problem-solving skills. For systems that need cognitive support or feedback when users lack basic knowledge, this point is crucial [14]. It involves evaluators simulating users' cognitive processes when thinking about the actions they took to accomplish tasks that based on their background knowledge. It is important for evaluators to put themselves in the user's shoes in order to produce good results [15]. The assessor evaluates the user interface and assesses how simple each step is for new [10].

Firstly, we identified the applications to evaluate and assess usability. Then we identified the tasks and users, who are actually evaluators, and determine the sequence of actions that user will take to carry out the task. After it, evaluators conducted the walkthrough and answer the following four questions after each step of task. These questions are aids to stimulating the user's cognitive process.

1. Will the user be trying to achieve the right effect?
2. Will the user discover that the correct action is available?
3. Will the user associate the correct action with the desired effect?
4. If the correct action is performed, will the user see that progress is being made?

In response to these questions, the user will answer a YES or NO along with reasons why their action was successful or unsuccessful. To further evaluate the main usability attributes which are effectiveness (accuracy of predictions), efficiency (time taken by AI model to give results), satisfaction of user and ease of use, users will be posed to multiple questions after the completion of each task. In response of these questions, users will describe the usability problems found and their recommendations.

IV. AI BASED M-HEALTH APPLICATIONS

We have selected three mHealth applications that uses Artificial Intelligence. These applications include, Ada, Babylon, and Ornament Health application. The criteria of selecting applications is based on use of AI model, availability on android and iOS, and diversity of applications. All the 3 applications use AI to predict disease from symptoms, personal recommendations etc., and available on Android and iOS.

Ada Health [31] is a mHealth application that uses AI algorithms to provide users with personalized symptom-checking and health information (Figure .2). In this application, users can identify potential health concerns and make informed decisions about seeking medical treatment. Users can input information about their symptoms, medical history, and other relevant health information into the application. A personalized report is generated based on this information, which suggests possible causes of the symptoms and recommends seeking medical care when necessary.

Babylon Health [32] is a another mHealth application that offers telehealth services to users. Home Screen of this application is shown in Figure 3. A virtual consultation can be scheduled with a healthcare professional, such as a doctor, nurse, or therapist, through the application. Medical records can also be accessed and prescriptions can be requested using the application. Specifically designed for non-emergency medical issues, Babylon Health provides users with convenient access to healthcare services. Using AI algorithms, the application guides patients to the most appropriate healthcare provider based on their symptoms and medical history. Ornament Health [33] is also a mHealth application that focuses on wellness and helping users achieve their health goals (Figure. 4). Users can track their physical activity, nutrition and other health metrics. Application's Ai models generated personalized recommendations for users using this collected information to help them improve their well-being and health. Users can also access wellness coaches through Ornament Health, who can answer questions and provide guidance on living a healthy lifestyle.

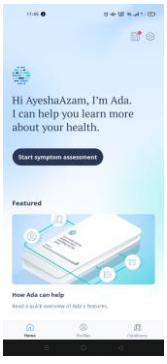

Figure 2. Ada Health Mobile App Home Screen

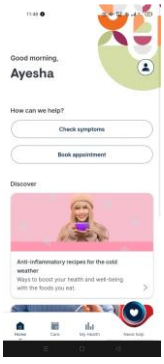

Figure 3. Babylon Health Mobile App Home Screen

Fugire 4. Ornament Health Mobile App Home Screen

A. Evaluators

For a cognitive walkthrough evaluation, a minimum of three evaluators is recommended to ensure that a variety of perspectives are represented [34]. It was found in a study by [35] that only three subjects are needed to uncover 65% of the problems, five are needed to uncover 80%, and nine are needed to uncover 95%. Our study was conducted by 15 evaluators. The evaluators include Ph.D students of which, 5 were developers, 5 were UX designers, 3 were the HCI experts, and 2 were the Ph.D. Scholars in Computer science. All of them have prior experience with mHealth applications. We aimed to enrich the results by bringing in different perspectives from individuals with different expertise, despite the higher cost associated with using more evaluators. This enabled us to evaluate the usability of mHealth applications in a comprehensive manner, which can lead to the design of more effective and user-friendly products.

V. DATA COLLECTION AND ANALYSIS METHOD

Each task was performed independently by evaluators on three applications in order to carry out the evaluation. If a problem arises afterward to achieve a task from a users' perspective, the evaluators could report back [13]. In addition to acting as an observer, the researcher, along with the evaluators, took notes on the data collection forms regarding questions, comments, and ambiguities relating to the evaluation process. To assess usability attributes, we present evaluators with a questionnaire following completion of each task. This questionnaire includes attributes are effectiveness, efficiency, ease of use, and satisfaction. The definition of these usability attributes are as follows:

1. Effectiveness: It refers to how well the AI model performs in accurately predicting a specific health outcome or disease diagnosis.
2. Efficiency: It refers to how quickly the application and user can perform a specific task or process.
3. Ease of Use: It refers to how easily and intuitively users can interact with the application to perform specific tasks or access relevant information.
4. Satisfaction: It refers to overall satisfaction that user get after using and experiencing the application.

The evaluators reviewed the details of the usability issues and made corrections or additions as needed after the evaluation process was complete. All the identified problems were added to the list of problems, while the repeated problems were removed from the list. To calculate the severity of each problem we used the (1) [36],

$$\text{Severity} = \text{Frequency} * \text{Impact} \quad (1)$$

Where, frequency is the number of times a problem is occurring and impact is severity of consequences of the problem. Symptom Checking given impact value 5, virtual consultation, personal recommendations, health tracking and health information were given 4, 3, 2, and 1 respectively.

 ISBN: 978-1-68558-096-4

VI. RESULTS AND DISCUSSION

An evaluation of three AI-based mHealth applications, including Ada health, Babylon health, and Ornament health, was conducted using Cognitive walkthrough method. The five tasks were performed by 15 evaluators to identify the problems with the application's usability. To further assess the usability of AI based tasks, we designed and presented a questionnaire to evaluators following the cognitive walkthrough of each task. In total, 56 problems were detected, of which 40 unique ones remained (14 from Ada health, 14 from Babylon and 12 from Ornament health application) after eliminating duplicates and combining the problems. Table 2 summarizes the unique problems identified and recommendations based on usability attributes.

TABLE II. A USABILITY ATTRIBUTE-BASED IDENTIFIED PROBLEMS AND RECOMMENDATIONS

Usability Attributes	Application	Identified Problem	Recommendations	Task
Effectiveness	Ada	Not relevant articles	Add articles relevant to history of patient.	Health Information
		Sometimes system provides a list of irrelevant diseases.	Re train the AI models	Symptom Checking
		AI model only enlist few possible diseases and nothing else.	Provide with medical treatment options as well.	Symptom Checking
	Babylon	No personal recommendation according to expectation	Retrain the model with updated data and new algorithms Or use collaborative filtering techniques	Personal Recommendations
		No option to search for articles of user's choice	Add search option so user can search for relevant information	Health Information
		Asks for driver license and other unnecessary details before booking consultation session.	Do not ask user for passport or driving license information.	Virtual Consultation
		AI model only enlist few possible diseases and nothing else.	Provide with medical treatment options as well.	Symptom Checking
	Ornament	limited options (disease) were given on "Ask Doctor page"	Add more variety of diseases to choose from	Health Tracking
		Presented data and statistics is difficult to understand.	Use user friendly language and visualization methods.	Health Information
Efficiency	Ada	No search option make user took a lot of time in order to search and get desired information.	Add search functionality	Health Information
	Babylon	No search option make user took a lot of time in order to search and get desired information.	Add search functionality	Health Information
		Duplicate buttons	Remove duplicated buttons of book appointment and symptom check	Virtual Consultation, Symptom Checking
	Ornament	Application give response to almost every touch after some time	Optimize the code to improve speed and responsiveness of application	Health information, Health Tracking, Personal recommendations
Ease of Use	Ada	Menus and options were not organized in logical manner	Restructure and group the menus and options in a more intuitive and user-friendly manner	Symptom checking
		It was not easy to find the information that we are looking for.	Categorize the data, Add search functionality on home page	Health Information, Symptom checking

		No "Go back" to home button	Add "Go back" button on every screen	Health Information
	Babylon	Presentation and design of application was not pleasant.	Use more pleasing color scheme and design elements	Symptom Checking, Health Information, Health Tracking, Virtual Consultation
		Navigation is tiresome.	Provide clear and concise labels for navigation. Optimize layout design for easy navigation.	Symptom Checking, Health Information, Health Tracking
	Ornament	Some evaluator reported that UI of application is not pleasing.	Use visually appealing color scheme, typography, and layout	Health Tracking
		Application gets stuck on Insight's page sometimes	Optimize the page loading and processing times by reducing unnecessary data from page	Health Tracking
		Restriction on Must select 3 topics to get insights on.	Remove this restriction	Health Information
		Navigating Back does not work on some pages	Make it work on every page.	Health Information, Health Tracking
Satisfaction	Ada	Sometimes results are not what user expected.	Improve the accuracy of the symptom checker algorithm by incorporating more comprehensive and up-to-date knowledge	Symptom checking
		Did not feel informed about health after using it	Provide comprehensive and personalized health information.	Health Information, Personal recommendations
		Some icons are different from their functions.	Use icons that have clear meanings for users.	Health Information, Personal recommendations
	Babylon	Very few articles to read	Add more user health history related articles.	Health Information
	Ornament	Due to time lagging, the most of the users are not very satisfied with application.	Work on improving speed and responsiveness.	Health Information, Personal recommendations, Health tracking

TABLE III. AVERAGE SEVERITY OF IDENTIFIED PROBLEMS AND AVERAGE TIME TAKEN BY EACH TASK

Tasks	**Severity in ADA**	**Severity in Babylon**	**Severity in Ornament**	**Average Time Taken in Ada (min)**	**Average Time Taken in Babylon (min)**	**Average Time Taken in Ornament (min)**
Symptom Checking	25	20	0	03:35	02: 58	00:00
Virtual Consultation	0	8	0	00:00	00:00	00:00
Personal Recommendations	6	3	6	03:50	04: 38	05:34
Health Tracking	0	4	10	00:00	02:10	04:23
Health Information	6	5	5	02: 35	01:25	03:59
Average	**7.4**	**8.0**	**4.2**	**03:20**	**02:48**	**04:39**

One of the problems that evaluators reported frequently is inaccurate and unexpected results of symptom checker model of applications and it is related to effectiveness of AI models and applications. Other research studies have found that health applications often suffer from poor accuracy, which can undermine their effectiveness [37], [38]. Patients are the most significant recipients and users of AI based mobile applications, thus ensuring its use in healthcare does not harm but rather benefits them should be a priority [39]. Hence, AI systems should be efficient and effective to provide user satisfaction [8]. AI in mobile health applications is a helpful tool [40]. AI has the potential to engage their users and develop significant and healthy

connections with them over time. Using advance machine and deep learning methods and large amount of data, the effectiveness, accuracy of algorithms can improve [41].

Another reoccurring problem that evaluators reported is no search bar (Figure. 5) when performing health information task, thus, affecting the efficiency usability attribute. Usability issues related to efficiency have also been identified in the literature [42], [43] reported frustration in users due to long waits. Hence, it is the need for applications to be faster and more responsive. Similarly, in another it was found to be frustrating and time-consuming for users to manually browse through a lot of information on health information pages without a search bar on interface [44]. A search bar added on page of an application makes it more efficient and enhances the user experience. A study by the Nielsen Norman Group [45] has shown that the presence of a search bar improves the efficiency of an application by allowing users to quickly find what they are looking for. Participants were asked to perform a series of tasks on a website, some with a search bar and some without. The results showed that participants were able to complete tasks faster when a search bar was present hence more efficient. To get health information, ornament restrict their user to select minimum of 3 topics to get insights on as shown in Figure. 6, ultimately affecting the ease-of-use usability attribute of application.

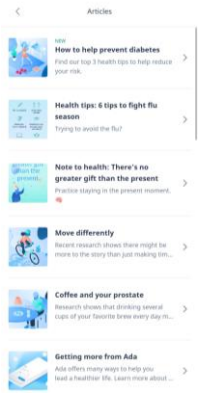

Figure 5. No search Bar on the Health Information Page

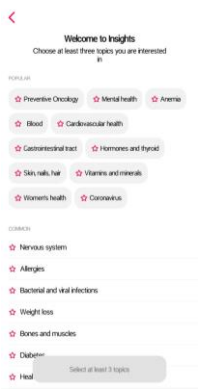

Figure 6. Restriction to select Minimum of 3 items to get insights on

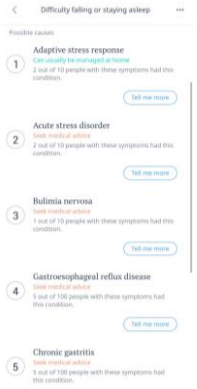

Figure 7. List of Possible diseases

Participants also reported that navigation was tiresome in Babylon. And in Ada Health app, Elements were not properly organized, making it difficult for them to look for the information they require. Go back button was also missing on some screens. However, AI applications need system behavior to be presented in a clear manner. Furthermore, despite dynamic system behavior, high consistency levels are achieved. The navigation design of an application impacts how easily the user can operate it [46].

The health applications Babylon and Ada do not offer personalized recommendations to users. Upon entering symptoms, these applications only generate a list of possible diseases as shown in Figure. 7. These symptoms are added to the user's history. However, evaluators have reported that they did not observe any articles personalized to their history that met their expectations. There are studies that identified this problem the lack of personalization in AI based mobile health applications. However, personalized recommendations are particularly important in AI based Healthcare applications, as they enable to provide personalized information to user to meet their expectations which ultimately affect the satisfaction of user [47]. Personalized recommendations significantly increased people's likelihood of adopting healthy lifestyle behaviors, according to a study [48]. Another study, shows that the personalized diabetes management recommendations generated by an AI algorithm improve glycemic control better [49]. Additionally, a study found that the lack of user-specific data is a major challenge in developing personalized mobile health applications [50]. Furthermore, a study concluded that personalization is crucial in AI systems used for clinical decision-making, as it allows for more accurate and effective diagnoses and treatment plans [51]. Health care AI apps have predominantly focused on AI's analytical capabilities, and data handling, but have neglected human factors perspectives, resulting in poorly designed apps [52]. Issues such as the need for simpler navigation and better design have been noted which affect the ease-of-use usability attribute [53]. Research has also high-lighted the importance of satisfaction in health applications, as a factor to determine the success of a health care facility [54].

Finally, the evaluators from our study recommended to retrain the models with updated data to provide personal recommendations that meet individual needs of the users, and results in overall good user experience. To address these issues, evaluators proposed following solutions. They suggested that the addition of a search bar in such applications can significantly enhance a user's experience since it makes it easier for them to find the desired information without having to navigate through multiple screens. Additionally, prior research has shown that search functionality in health applications is extremely important [55],[56]. To address the problem of inaccurate predictions, it is recommended that these applications provide users with more personalized recommendations. It is suggested that a more relevant and accurate recommendation can be provided by collecting more user data and incorporating it into the AI algorithms. This suggestion is also reinforced by previous research that emphasizes the importance of personalization and personalized recommendations in mHealth applications [57]–[60]. Users who are unfamiliar with the application may be confused and frustrated by the absence of a "navigate to back" button on certain screens. By adding this feature to all screens, the application's overall usability can be improved and user frustration can be reduced. Additionally, it is recommended that these applications should include more engaging and attractive functionalities, such personalized feedback, goal-setting, and performance reporting, to improve the overall user experience. These features can increase application commitment and user motivation, which will lead to improved health results.

 ISBN: 978-1-68558-096-4

VII. CONCLUSION

In conclusion, the widespread use of mobile healthcare application with making use of artificial intelligence technology provide numerous healthcare services to their users. However, as the number of mobile applications in increasing day by day, evaluating their usability in terms of effectiveness of AI systems they provide, efficiency in terms of time user take to perform a task, ease of using application and over experience of user is crucial factor in their success. Cognitive walkthrough is one of many methods of usability evaluation. In this study, we selected 3 applications that make use of AI in their features, on the basis of their popularity, and availability to evaluate their usability through cognitive walkthrough. The identified tasks are symptom checking, virtual consultation, personal recommendations, health tracking and health information. 27 unique problems were identified after eliminating the repeating ones. The most of the problems were reported in symptom checking and health information tasks, 9 and 16 respectively. Since health information and health tracking impact value are lesser than symptom checking and virtual consultation. Therefore, the average severity of problems in Ada, Babylon and Ornament are 7.4, 8.0 and 4.2 respectively. Babylon has the most severity due to the high impact of symptom checking and virtual consultation tasks. The average time taken by users in these applications is varied, ornament being taking the longest time to complete tasks, because evaluators reported unresponsiveness issues in Ornament application. The evaluators provided recommendations for the identified problems to improve the effectiveness, efficiency, ease of use, and satisfaction of these applications. From our study, it is clear that evaluating the usability during the development and design of mHealth applications, especially those that use AI-based features is crucial to ensure the success and effectiveness of application. Adding more usability evaluation methods can enrich such studies taking into account also additional mobile applications. Involving larger number of users can be seen as an extension of this research project.

AUTHORS CONTRIBUTION

The author has formalized the idea, conducted a comprehensive literature review and then conducted the evaluation. The author then completed data analysis, discussion and writing up the paper.

ACKNOWLEDGMENT

This research was funded by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, grant number 523.

REFERENCES

[1] A. Muro-Culebras et al., “Tools for Evaluating the Content, Efficacy, and Usability of Mobile Health Apps According to the Consensus-Based Standards for the Selection of Health Measurement Instruments: Systematic Review,” JMIR Mhealth Uhealth 2021;9(12):e15433 https://mhealth.jmir.org/2021/12/e15433, vol. 9, no. 12, p. e15433, Dec. 2021, doi: 10.2196/15433.

[2] S. S. E. Alwi, M. A. A. Murad, S. Abdullah, and A. Kamaruddin, “A Prototype Development and Usability Evaluation of an E-health System,” Proceedings - AiIC 2022: 2022 Applied Informatics International Conference: Digital Innovation in Applied Informatics during the Pandemic, pp. 34–39, 2022, doi: 10.1109/AIIC54368.2022.9914606.

[3] “mHealth Apps Market Size, Share & Trends Analysis Report by Type (Fitness, Medical), by Region (North America, Europe, Asia Pacific, Latin America, Middle East & Africa), and Segment Forecasts, 2022-2030.” https://www.researchandmarkets.com/reports/4396364/mhealth-apps-market-size-share-and-trends (accessed Feb. 22, 2023).

[4] D. E. Jake-Schoffman et al., “Methods for Evaluating the Content, Usability, and Efficacy of Commercial Mobile Health Apps,” JMIR Mhealth Uhealth 2017;5(12):e190 https://mhealth.jmir.org/2017/12/e190, vol. 5, no. 12, p. e8758, Dec. 2017, doi: 10.2196/MHEALTH.8758.

[5] R. Alturki, V. G.-J. formative research, and undefined 2019, “The development of an Arabic weight-loss app Akser Waznk: qualitative results,” formative.jmir.org, Accessed: Feb. 22, 2023. [Online]. Available: https://formative.jmir.org/2019/1/e11785/

[6] A. I. Alharbi, V. Gay, M. J. Alghamdi, R. Alturki, and H. J. Alyamani, “Towards an Application Helping to Minimize Medication Error Rate,” Mobile Information Systems, vol. 2021, 2021, doi: 10.1155/2021/9221005.

[7] K. Komalavalli and R. Hemalatha, S. D.-S. I. Journal, and undefined 2020, “A Survey of Artificial Intelligence in Smart Phones and Its Applications among the Students of Higher Education in and around Chennai City.,” ERIC, doi: 10.34293/education.v8i3.2379.

[8] C. Cutillo, K. Sharma, L. Foschini, … S. K.-N. digital, and undefined 2020, “Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency,” nature.com, Accessed: Mar. 10, 2023. [Online]. Available: https://www.nature.com/articles/s41746-020-0254-2

[9] R. Khajouei and A. Ameri, Y. J.-I. journal of medical informatics, and undefined 2018, “Evaluating the agreement of users with usability problems identified by heuristic evaluation,” Elsevier, Accessed: Feb. 22, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505618301254

[10] C. Lewis, P. Polson, C. Wharton, and J. Rieman, “Testing a Walkthrough Methodology for Theory-Based Design of Walk-Up-and-Use Interfaces”.

[11] C. Lewis, C. Wharton. of human-computer interaction, and undefined 1997, “Cognitive walkthroughs,” Elsevier, Accessed: Feb. 22, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444818621500960

[12] N. Fasihah Jamaludin, “The usability evaluation of adolescent depression screening model using cognitive walkthrough

[13] ,” Asia-Pacific Journal of Information Technology and Multimedia Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik, vol. 11, no. 2, pp. 71–88, 2022, doi: 10.17576/apjitm-2022-1102-06.

[14] M. Farzandipour, E. Nabovati, H. Tadayon, and M. Sadeqi Jabali, “Usability evaluation of a nursing information system by applying cognitive walkthrough method,” Int J Med

Inform, vol. 152, p. 104459, Aug. 2021, doi: 10.1016/J.IJMEDINF.2021.104459.

[15] M. Georgsson, N. Staggers, E. Årsand, and A. Kushniruk, "Employing a user-centered cognitive walkthrough to evaluate a mHealth diabetes self-management application: A case study and beginning method validation," J Biomed Inform, vol. 91, p. 103110, Mar. 2019, doi: 10.1016/J.JBI.2019.103110.

[16] L. O. Bligård and A. L. Osvalder, "Enhanced cognitive walkthrough: Development of the cognitive walkthrough method to better predict, identify, and present usability problems," Advances in Human-Computer Interaction, vol. 2013, 2013, doi: 10.1155/2013/931698.

[17] M. Maguire.-I. journal of human-computer studies and undefined 2001, "Methods to support human-centred design," Elsevier, vol. 55, pp. 587–634, 2001, doi: 10.1006/ijhc.2001.0503.

[18] "Fifth Annual 'Pulse of Online Health' Survey Finds 66% of Americans Eager To Leverage Digital Tools To Manage Personal Health." https://www.prnewswire.com/news-releases/fifth-annual-pulse-of-online-health-survey-finds-66-of-americans-eager-to-leverage-digital-tools-to-manage-personal-health-300039986.html (accessed Jun. 13, 2023).

[19] A. Roess . J. Med and undefined 2017, "The promise, growth, and reality of mobile health-another data-free zone," researchgate.net, doi: 10.1056/NEJMp1713180.

[20] H. Cho, P. Y. Yen, D. Dowding, J. A. Merrill, and R. Schnall, "A multi-level usability evaluation of mobile health applications: A case study," J Biomed Inform, vol. 86, pp. 79–89, Oct. 2018, doi: 10.1016/J.JBI.2018.08.012.

[21] R. Schnall, J. Mosley, … S. I.-J. mHealth and, and undefined 2015, "Comparison of a user-centered design, self-management app to existing mHealth apps for persons living with HIV," mhealth.jmir.org, Accessed: Feb. 21, 2023. [Online]. Available: https://mhealth.jmir.org/2015/3/e91/

[22] O. V. Bitkina, H. K. Kim, and J. Park, "Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges," Int J Ind Ergon, vol. 76, p. 102932, Mar. 2020, doi: 10.1016/J.ERGON.2020.102932.

[23] Y. A. Daraghmi, B. Yahya, and Y. Daraghmi, " Has Covid-19 affected software usability: mobile accounting system as a case" J Theor Appl Inf Technol, vol. 31, no. 2, 2023, Accessed: Feb. 17, 2023. [Online]. Available: www.jatit.org

[24] T. Krause, E. Jolkver, S. Bruchhaus, P. M. Kevitt, M. Kramer, and M. Hemmje, "A Preliminary Evaluation of “GenDAI”, an AI-Assisted Laboratory Diagnostics Solution for Genomic Applications," BioMedInformatics 2022, Vol. 2, Pages 332-344, vol. 2, no. 2, pp. 332–344, Jun. 2022, doi: 10.3390/BIOMEDINFORMATICS2020021.

[25] A. S. Dahri, A. S. Dahri, A. Al-Athwari, and A. Hussain, Usability Evaluation of Mobile Health Application from AI Perspective in Rural Areas of Pakistan, vol. 14, no. 15. International Association of Online Engineering, 2019. doi: 10.3991/ijim.v13i11.11513.

[26] M. N. Islam, S. R. Khan, N. N. Islam, M. Rezwan-A-Rownok, S. R. Zaman, and S. R. Zaman, "A Mobile Application for Mental Health Care During COVID-19 Pandemic: Development and Usability Evaluation with System Usability Scale," Advances in Intelligent Systems and Computing, vol. 1321, pp. 33–42, 2021, doi: 10.1007/978-3-030-68133-3_4.

[27] N. A. Zaini, S. F. M. Noor, and T. S. M. T. Wook, "Evaluation of APi interface design by applying cognitive walkthrough," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, pp. 306–315, 2019, doi: 10.14569/IJACSA.2019.0100241.

[28] M. Farzandipour, E. Nabovati, and M. S. Jabali, "Comparison Usability Evaluation Methods in a Health Information System: Heuristic Evaluation versus Cognitive Walkthrough Method," Oct. 2021, doi: 10.21203/RS.3.RS-871961/V1.

[29] T. Kim et al., "Assessing the usability of a prototype emergency medicine patient-centered electronic health record display," Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, pp. 424–425, Jul. 2018, doi: 10.1109/ICHI.2018.00083.

[30] P. G. Poison and C. H. Lewis, "Theory-Based Design for Easily Learned Interfaces," Hum Comput Interact, vol. 5, no. 2–3, pp. 191–220, 1990, doi: 10.1080/07370024.1990.9667154.

[31] P. Polson, C. Lewis, J. Rieman, C. W.-I. J. of man, and undefined 1992, "Cognitive walkthroughs: a method for theory-based evaluation of user interfaces," Elsevier, Accessed: Mar. 01, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002073739290039N

[32] "Health. Powered by Ada." https://ada.com/ (accessed Mar. 01, 2023).

[33] "Babylon Health UK - The Online Doctor and… | Babylon Health." https://www.babylonhealth.com/en-gb (accessed Mar. 01, 2023).

[34] "Homepage | Ornament." https://ornament.health/ (accessed Mar. 01, 2023).

[35] R. Khajouei, M. Zahiri Esfahani, and Y. Jahani, "Comparison of heuristic and cognitive walkthrough usability evaluation methods for evaluating health information systems," J Am Med Inform Assoc, vol. 24, no. e1, pp. e55–e60, Apr. 2017, doi: 10.1093/JAMIA/OCW100.

[36] R. A. Virzi, "Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?," https://doi.org/10.1177/001872089203400407, vol. 34, no. 4, pp. 457–468, Nov. 2016, doi: 10.1177/001872089203400407.

[37] R. S. Pressman, "Software engineering : a practitioner's approach," p. 941.

[38] S. Jayakumar et al., "Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study," NPJ Digit Med, vol. 5, no. 1, Dec. 2022, doi: 10.1038/S41746-021-00544-Y.

[39] M. Modiba, "Artificial intelligence for the improvement of records management activities at the Council for Scientific and Industrial Research," Journal of the South African Society of Archivists, vol. 55, pp. 16–26, Nov. 2022, doi: 10.4314/jsasa.v55i.2.

[40] P. Esmaeilzadeh, "Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives," BMC Med Inform Decis Mak, vol. 20, no. 1, Jul. 2020, doi: 10.1186/S12911-020-01191-1.

[41] S. Berrouiguet, M. L. Barrigón, J. L. Castroman, P. Courtet, A. Artés-Rodríguez, and E. Baca-García, "Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: The Smartcrises study protocol," BMC Psychiatry, vol. 19, no. 1, Sep. 2019, doi: 10.1186/S12888-019-2260-Y.

[42] A. S. Alzahrani, V. Gay, R. Alturki, and M. J. Alghamdi, "Towards Understanding the Usability Attributes of AI-Enabled eHealth Mobile Applications," J Healthc Eng, vol. 2021, 2021, doi: 10.1155/2021/5313027.

[43] D. Li, "5G and intelligence medicine—how the next generation of wireless technology will reconstruct healthcare?," Precis Clin Med, vol. 2, no. 4, pp. 205–208, Dec. 2019, doi: 10.1093/PCMEDI/PBZ020.

[44] O. Strachna and O. Asan, "Systems Thinking Approach to an Artificial Intelligence Reality within Healthcare: From Hype to Value," ISSE 2021 - 7th IEEE International Symposium on

Systems Engineering, Proceedings, Sep. 2021, doi: 10.1109/ISSE51541.2021.9582546.

[45] "Importance of Adding Search Bar to your Website - The Next Scoop." https://thenextscoop.com/add-search-bar-to-website/ (accessed Mar. 10, 2023).

[46] "Search: Visible and Simple." https://www.nngroup.com/articles/search-visible-and-simple/ (accessed Mar. 10, 2023).

[47] L. Wiebelitz, P. Schmid, T. Maier, and M. Volkwein, "Designing User-friendly Medical AI Applications - Methodical Development of User-centered Design Guidelines," Proceedings - 2022 IEEE International Conference on Digital Health, ICDH 2022, pp. 23–28, 2022, doi: 10.1109/ICDH55609.2022.00011.

[48] "4 Reasons Healthcare should use Personalisation | Codehouse." https://www.codehousegroup.com/insight-and-inspiration/digital-strategy/4-reasons-healthcare-should-use-personalisation (accessed Mar. 10, 2023).

[49] S. M. Kelders, R. N. Kok, H. C. Ossebaard, and J. E. W. C. Van Gemert-Pijnen, "Persuasive System Design Does Matter: A Systematic Review of Adherence to Web-Based Interventions," J Med Internet Res 2012;14(6):e152 https://www.jmir.org/2012/6/e152, vol. 14, no. 6, p. e2104, Nov. 2012, doi: 10.2196/JMIR.2104.

[50] D. Bertsimas, N. Kallus, A. M. Weinstein, and Y. D. Zhuo, "Personalized Diabetes Management Using Electronic Medical Records," Diabetes Care, vol. 40, no. 2, pp. 210–217, Feb. 2017, doi: 10.2337/DC16-0826.

[51] J. ANDRES. "Exploring AI-based personalization of a mobile health intervention and its effects on behavior change, motivation, and adherence," 2021. http://reports-archive.adm.cs.cmu.edu/anon/hcii/CMU-HCII-21-104.pdf (accessed Mar. 20, 2023).

[52] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science (1979), vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/SCIENCE.AAX2342/SUPPL_FILE/AAX2342_OBERMEYER_SM.PDF.

[53] "'Most healthcare apps not up to NHS standards' - BBC News." https://www.bbc.com/news/technology-56083231 (accessed Mar. 20, 2023).

[54] B. H. Kann, A. Hosny, and H. J. W. L. Aerts, "Artificial intelligence for clinical oncology," Cancer Cell, vol. 39, no. 7, pp. 916–927, Jul. 2021, doi: 10.1016/J.CCELL.2021.04.002.

[55] F. Manzoor, L. Wei, A. Hussain, M. Asif, and S. I. A. Shah, "Patient Satisfaction with Health Care Services; An Application of Physician's Behavior as a Moderator," Int J Environ Res Public Health, vol. 16, no. 18, Sep. 2019, doi: 10.3390/IJERPH16183318.

[56] W. J. Gordon, A. Landman, H. Zhang, and D. W. Bates, "Beyond validation: getting health apps into clinical practice," NPJ Digit Med, vol. 3, no. 1, Dec. 2020, doi: 10.1038/S41746-019-0212-Z.

[57] C. Lee Ventola, "Mobile Devices and Apps for Health Care Professionals: Uses and Benefits," Pharmacy and Therapeutics, vol. 39, no. 5, p. 356, 2014, Accessed: Mar. 16, 2023. [Online]. Available: /pmc/articles/PMC4029126/

[58] E. M. Grua, M. De Sanctis, I. Malavolta, M. Hoogendoorn, and P. Lago, "An evaluation of the effectiveness of personalization and self-adaptation for e-Health apps," Inf Softw Technol, vol. 146, p. 106841, Jun. 2022, doi: 10.1016/J.INFSOF.2022.106841.

[59] U. Josefsson, "Association for Information Systems AIS Electronic Library (AISeL) Exploring e-patients ' heterogeneity: towards personalized e-health applications," p. 2006, Accessed: Mar. 16, 2023. [Online]. Available: http://aisel.aisnet.org/ecis2006

[60] A. Saad, H. Fouad, and A. A. Mohamed, "Situation-aware recommendation system for personalized healthcare applications," J Ambient Intell Humaniz Comput, vol. 1, pp. 1–15, Feb. 2021, doi: 10.1007/S12652-021-02927-1/TABLES/4.

[61] X. Zhou, W. Liang, K. I. K. Wang, and S. Shimizu, "Multi-Modality Behavioral Influence Analysis for Personalized Recommendations in Health Social Media Environment," IEEE Trans Comput Soc Syst, vol. 6, no. 5, pp. 888–897, Oct. 2019, doi: 10.1109/TCSS.2019.2918285.

 ISBN: 978-1-68558-096-4

Leveraging Digital Twins for Condition Monitoring in Railway Infrastructure

Lucas Rocha
Department of Informatics Engineering
Faculty of Engineering of the University of Porto
Porto, Portugal
up201902814@up.pt

Gil Gonçalves
Department of Informatics Engineering
Faculty of Engineering of the University of Porto
Porto, Portugal
gil@fe.up.pt

***Abstract*—Rail transport services are emerging as a major sustainable transportation option, especially in metropolitan areas. However, these services are significantly dependent on high investments and complex logistics, which creates the need to identify opportunities to minimize the waste of resources so that these services remain affordable and competitive. The digital twin, one of the core concepts of the Industry 4.0 paradigm, enables detailed, real-time monitoring of the state of a piece of equipment during its operation. For this reason, the digital twin represents an important support in ensuring sound decision making. This work seeks to explore the potential of the digital twin to support the monitoring of the conditions of railway vehicles and railroad infrastructure. First, a research was conducted on the implementations and studies of digital twins carried out in recent years. Next, two digital twin prototypes – a digital twin of a railway vehicle model, and another one for a section of a railroad - were developed. Both prototypes are composed of a relational database for storing the data on the operational conditions of the equipment, a mobile application that works as a dashboard of the digital between the application and the database. With the developed prototypes, it was possible to glimpse how the digital twin concept provides a deeper knowledge of the working conditions of the components of a train. Besides, the prototype also supports preventive maintenance through the analysis of the historical evolution of the data collected from these components. The results of this study also allow the identification of possible improvements and research opportunities for future work.**

Keywords- rail transport; digital twin; monitoring; Industry 4.0.

I. INTRODUCTION

Accelerated technological and scientific progress over the last few centuries has enabled the worldwide growth of industrialization [1]. This phenomenon began with the First Industrial Revolution at the end of the 18th century, which introduced mechanical manufacturing facilities powered by water and steam, as well as equipment such as the mechanical loom. A century later, the Second Industrial Revolution was marked by the spread of mass-production factories powered by electricity, along with the introduction of division of labor. The Third Industrial Revolution, which took place in the early 1970s, was characterized by the employment of electronics and information technology (IT) to achieve an even greater level of automation in manufacturing processes [2]. The emergence of Internet of Things (IoT) technologies has led to a new industrial paradigm shift, which has been dubbed Industry 4.0 [3]. The transition from centralized industrial control systems to decentralized intelligent systems can be considered the central principle behind the Industry 4.0 paradigm. The Industrial Internet of Things (IIoT) – which refers to the real time collection and sharing of data between products, components and industrial machines – allows industrial systems to adapt their behavior to different operating conditions [3].

One of the key technologies in the Industry 4.0 paradigm is the digital twin concept. The digital twin makes it possible to integrate the physical world with the virtual world by creating a virtual representation of a piece of equipment or object. For this reason, a digital twin facilitates the building of information systems that offer a solid basis for decision-making [4]. Consequently, this concept has been adopted not only in the context of industrial production, but also in sectors such as urban planning and health [5]. The digital twin is also a promising technology for improving and modernizing rail transport processes. As these operations are significantly dependent on high investments and complex logistics, it is desirable to optimize costs related to equipment maintenance and modernization, as well as minimize downtime. The digital twin concept can be employed to tackle these challenges, as it enables the monitoring of the operational conditions of rail transport systems and its resources [6].

Rail transport represents a crucial service for a country's logistical and economic scenario, as it allows the transportation of a large number of passengers and heavy loads over long distances. In addition, it is an important means of transportation for maintaining a sustainable economy [7]. Considering the worsening climate crisis - to which the saturation of the road system contributes significantly - it is necessary to increase the competitiveness and attractiveness of rail transport services. For this reason, the European Commission has strongly recommended an increase in the share of rail transport compared to road transport [7].

The present paper seeks to explore the potential of the digital twin concept for rail transport support, especially regarding the monitoring of the conditions of railway vehicles and railroads, and the implementation of predictive maintenance. To this end, two digital twin prototypes were

 ISBN: 978-1-68558-096-4

developed: a digital twin of a train model, as well as a digital twin of a section of railroad. Both prototypes consist of a database, a web server and a mobile application for data visualization. This work was carried out as part of the Ferrovia 4.0 research project and was accompanied by project partners throughout the development process.

The remainder of this work is structured as follows. Section II details related work in digital twins. Section III describes the methods and tools used in the development of the present work. Section IV delineates the process of implementation of the digital twin prototypes. The employed evaluation method is explained in Section V, and its results are presented and discussed in Section VI. Lastly, Section VII presents the conclusions and opportunities for future work.

II. Related Work

The concept, as well as the term "digital twin" itself, were introduced in 2003 by Grieves in the Product Lifecycle Management course at the University of Michigan. At the time, the notion of virtual representations of physical products was in its infancy, and the technological limitations meant that data on the real product had to be collected manually through paper [8]. These same limitations were the main cause behind the lack of practical studies related to digital twins in the years following its introduction [9].

Although not very specific, a preliminary digital twin model was proposed by Grieves at the time. This model had three main components: the real product, the virtual product and the data connections responsible for linking the real and virtual products [8][9]. Accelerated advances in communication, sensor, simulation and big data technologies over the course of the 2000s have contributed to the rise in the number of digital twin studies over the last decade, as these advances enabled the automated collection of product data [9]. Fig. 1 shows the growth in the number of scientific publications related to digital twins since the introduction of this concept.

In 2011, the first scientific journal article on the concept of the digital twin was published [9]. In this article, Tuegel et al., [10] sought to apply the digital twin concept to improve the prediction of the useful life of aircrafts, which at the time consisted of using individual physical models of the different categories of stress exerted on the airframe. To this end, the authors proposed the use of high-fidelity models for each unit of a specific variety of aircraft in an inventory. By using data on the estimated flight path and expected maneuvers for a given mission assigned to the aircraft, these models could perform a simulation and calculate the level of stress that would be exerted on the machine's structure as a result of the flight [10].

In an article published in the following year, NASA formalized the definition of a digital twin as a multi-physics, multi-scale, probabilistic, high-fidelity simulation that uses historical data, sensor data and physical models to reflect the state of a real product. In this article, the authors proposed the use of digital twins to address the shortcomings of conventional vehicle certification and fleet management methods employed at NASA and the United States Air Force [11].

Guo et al., [12] proposed a modular approach to assist in the development of a flexible digital twin for evaluating factory designs. The authors make use of parameterized and reusable modules that correspond to real physical entities to make the process of developing the digital twin more flexible, dynamic and faster. This approach results in a simulation model made up of modules that are independent of each other, which speeds up any changes to the model and enables collaboration between multiple designers [12].

Vachalek et al., [13] presented a digital twin-based approach for production line optimization. The authors used a simulated pneumatic cylinder production line paired with a detailed digital twin of the actual physical process. In addition to enabling the simulation of alternative manufacturing scenarios by modifying production parameters, the digital twin was also able to monitor the process in real time and identify opportunities for minimizing resource consumption [13].

Tao and Zhang [14] proposed the concept of the digital twin shop-floor, which consists of a virtual reproduction of the geometry, behavior and rules of a given shop-floor. This digital twin is updated in real time according to data related to the operations carried out on the physical shop floor. This enables the digital twin to carry out simulation, evaluation and optimization tasks, as well as regulating physical operations automatically as required [14].

Haag and Anderl [15] presented a proof of concept of the digital twin by employing it to a bending beam test procedure. The authors designed a test bench in which two actuators are used to apply force to both sides of the beam in order to make it bend. Integrated sensors are responsible for measuring the resulting force and calculating the displacement of the beam by using the difference in the position of the actuators. This data is sent to a digital twin of the test bench, which consists of a three-dimensional model of the built structure and a dashboard. Using this dashboard, users can monitor data about the force applied to the beam and the degree of displacement, as well as control the test bench's actuators [15].

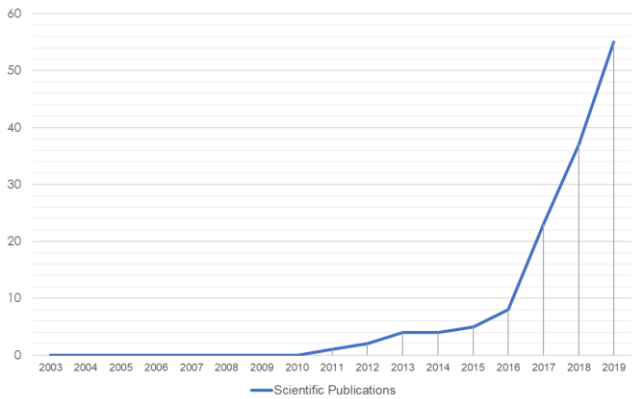

Figure 1. Approximated number of digital twin-related scientific publications per year. Adapted from [5][9]

Liu et al., [16] developed a digital twin-based approach to support the planning of diesel engine machining processes by analyzing and reusing knowledge from previous procedures. The digital twin built by the authors consists of geometry information and data on the current state of the equipment. This data was then combined with the accumulated process knowledge and then filtered through a similarity calculation algorithm, which discarded the process knowledge information that did not correspond to the current operation. The result was the set of process knowledge that was a candidate for the optimization procedure. Finally, the diesel engine components were applied to a prototype module in order to verify the effectiveness of the proposed method [16].

Jeschke and Grassmann [6] proposed a strategy for implementing digital twins in the German rail transport system. The authors presented a use case of an Intercity Express train, a high-speed rail transportation service that connects several cities in Germany and other European countries. According to the authors, by allowing the representation of a real object in a virtual environment, the digital twin concept enables the monitoring of the rolling stock in real time and the identification of unplanned changes in service operations. Through data-based evaluation and simulation, a digital twin can make early predictions of future events, which enables predictive maintenance and preventive measures against possible failures, as well as avoiding the waste of financial resources. The authors also identify some of the biggest obstacles to the implementation of digital twins in the context of the German rail transport service. They point to the absence of technical norms and standards for the interoperable operation of digital twins on a network, as well as legal barriers to obtaining and using data from pre-existing control and monitoring systems on the system's trains [6].

The Alstom and Simplan companies have developed a digital twin – by using the Anylogic simulation tool - with the aim of optimizing transport services on the West Coast Main Line, one of the UK's most important railways. Although there are fixed schedules for train operations, the need for maintenance regimes and the possibility of faults or accidents make it extremely difficult to predict the location of trains a few days in advance. Considering that simulations based on fixed data are inadequate in this case, a digital twin made it possible to explore different scenarios for optimizing rail services more efficiently and accurately [17].

III. MATERIALS AND METHODS

In order to achieve the desired objectives, two digital twin prototypes were built: the first one consists of a digital twin of an ARCO train, based on the vehicles operated by public transport services in Portugal. The second prototype is a digital twin of a 5km section of railroad based on a real section that is part of the Portuguese railway system. Both prototypes include mobile applications which serve as dashboards for the digital twins. These applications offer a visualization of metrics relating to a series of damage indicators, which are associated with the components of the vehicles and the railroad infrastructure.

The Unity game engine [18] was chosen tool for the development of the mobile applications. Although primarily designed for video game development, game engines are extremely versatile tools that offer a wide range of programming libraries and plugins for building interactive software. In addition, many of the main game engines on the market are either free for non-commercial use or are open-source projects. Unity was specifically chosen because for its versatility, its extensive support for 3D graphics and third-party plugins, as well as for being frequently employed in other research work on the topic of digital twins [5].

The data used by the digital twins was stored in relational databases, which are accessed by the mobile applications via requests to PHP files stored on an Apache web server. We decided to use relational databases as they allow for greater organization and structuring of the data. MySQL was selected as the relational database management system as the PHP language offers native support for the software.

We also chose to run both the databases and the web server inside Docker containers. Docker, unlike hypervisors, performs software virtualization at the operating system level, using individual user space instances called containers. Each instance contains an application and its dependencies and is completely isolated from other instances and the rest of the operating system. Fig. 2 illustrates the structure of the developed system.

Partners involved in the research project in which the proposed work is framed accompanied the development of the prototypes through presentations in meetings and workshops. The final validation of the prototypes was carried out through a quality assessment survey, which was sent to the project partners to assess both the degree of suitability of the graphic interface of the mobile applications, and the potential of the prototypes to support railway operations in a real context.

IV. IMPLEMENTATION

This section details the structure of the proposed prototypes, as well as the functionalities provided by each of digital twin's mobile applications.

A. Railway Vehicle Digital Twin

The mobile dashboard of the railway vehicle digital twin provides the user with a view of the metrics about two damage indicators associated with the components of the hypothetical vehicles: an indicator of the transmissibility of damage to the axle boxes, and an indicator of the length of wheel flats. The data relating to these indicators is synthetic and was obtained through simulations of different damage scenarios.

Through the mobile application, the user is able to select the specific vehicle and railway carriage they wish to analyze. The application interface also features a computer-aided design (CAD) model that represents an abstracted view of the railway carriage. This model was built using the Blender 3D modeling tool.

 ISBN: 978-1-68558-096-4

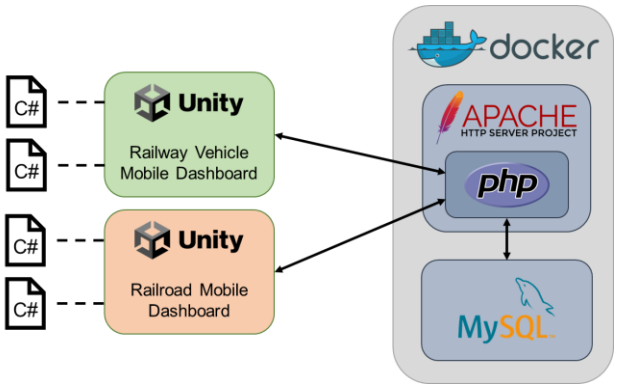

Figure 2. Structure of the developed digital twin system

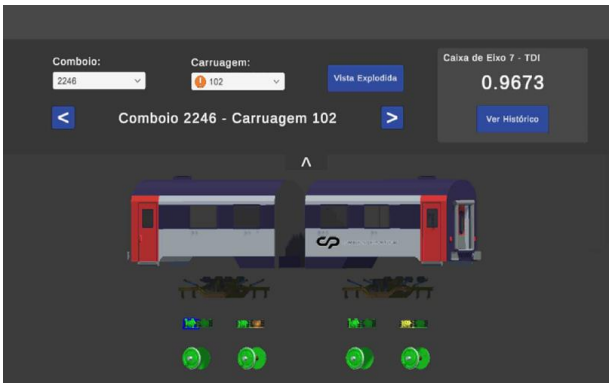

Figure 3. Main view of the railway vehicle's mobile dashboard

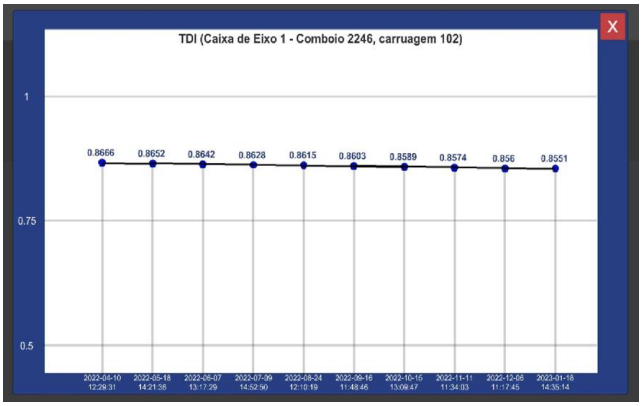

Figure 4. History graph of the damage indicator

Vehicle and railway carriage selection is carried out through dropdown menus. It is also possible to navigate along the carriages of a vehicle using two navigation buttons. The component of the vehicle is selected directly on the vehicle's CAD model by using touch input. As this prototype only includes damage indicators for axle boxes and wheelsets, these are only component types that can be selected by the user. An indication of the level of damage in each carriage is displayed next to their respective carriage option in the carriage selection dropdown menu. This indication has four different levels, which are dependent on the severity of the damage measured on a given carriage: a green icon, which corresponds to the absence of damage or the presence of superficial irregularities only; a yellow icon, which represents the presence of slight damage; an orange icon, which points to the existence of significant damage; or a red icon, which warns of the presence of serious damage. Similarly, each of the axle boxes and wheelsets in the CAD model are displayed in one of these colors, according to the level of damage shown by the indicators with which they are associated. The user interface also features an "Exploded View" button, which can be used by the user to switch the vehicles' view between the standard view - with the vehicle properly assembled - and the exploded view - in which the carbody, bogies, axleboxes and wheelsets are displayed as if they were disassembled.

An indication of the currently selected component, as well as the most recent measurement of the damage indicator to which the component is associated, are displayed in the left corner of the top menu. In addition, the user can also view a history graph of the indicator through the "View History" option. The user can navigate along the graph by tapping the left and right sides of the screen. Lastly, the user can also activate or deactivate the top menu freely. When it is deactivated, the CAD model of the vehicle takes up the entire screen. Fig. 3 shows the main view of the vehicle's mobile dashboard. Fig. 4 presents the history graph of the damage indicator.

B. Railroad Infrastructure Digital Twin

In a similar way to the railway vehicle's digital twin, the data on the railroad's damage indicators is synthetic and was obtained through simulations of several damage scenarios.

This data refers to vertical and horizontal irregularity indicators on both the left and right rails. Each point on a 5km stretch of a railroad has values associated with these indicators. For visualization purposes on the prototype, we decided to use a 100m interval between each point. Therefore, fifty points along this stretch were taken into account.

A static image obtained using the Google Street View API was associated with each of these points. The Uniform Resource Locator (URL) addresses of the images are stored in a MySQL database, as is the data on the indicators of irregularities in the infrastructure. The railway prototype's mobile application, similarly to how the vehicle prototype works, accesses the database by requesting PHP files and displays the static image associated with the selected point on the railroad. As Unity does not offer native support for displaying web pages, the third-party plugin UniWebView was employed to perform this task.

Through the mobile dashboard, the user is able to navigate along the fifty points of the railroad and interact with a map of the railroad section. This map was obtained via Google Maps and represents an aerial view of the railroad. Similarly to the approach used in the vehicle's digital twin, alert icons with different colors are displayed on the map according to the severity of the irregularities at each point. These alerts can be yellow, light orange, dark orange or red, which correspond to light, significant, very significant and serious damage respectively. By tapping on one of these

 ISBN: 978-1-68558-096-4

alerts, the user can find out what the maximum permitted speed is for the selected point, given the level of irregularity indicated. The user interface of the railroad's mobile dashboard is shown by Fig. 5.

V. EVALUATION

The evaluation of the proposed prototypes was done on meetings throughout the development of the work and through an online quality assessment survey sent to the Ferrovia 4.0 project partners, which served as a complement to the discussions raised at the meetings. The present paper will only discuss the evaluation of the railway vehicle prototype, as the railroad prototype will be assessed at a later date.

We decided to formulate the survey items using the Likert scale, which is a technique for measuring respondents' opinions and attitudes towards a series of statements that represent value judgments. The respondent must indicate their attitude towards each statement on an ascending scale of agreement. This scale usually has five values, which are represented by the numbers 1 to 5 [19]. This method was chosen because it is the most suitable technique for measuring partners' opinions on the quality of the mobile application's graphic interface and the potential of the prototype as a whole. In addition, the Likert scale is widely used for the assessment of software usability evaluation [20].

The survey was developed in an online format through the Google Forms tool and sent to partners via email. We opted for an online survey because it makes communication with partners easier and faster, as it allows participants to respond to the survey at any time. Alongside the survey, we also included a demonstration video of the railway vehicle's digital twin prototype, which respondents had to watch before submitting their answers. This was done to simplify the evaluation process and prevent partners from relying solely on the documentation made during the development of the project to answer the survey. A five-value Likert scale was used for the survey items, which are represented by the numbers 1 to 5 and interpreted in ascending order as "strongly disagree", "partly disagree", "neutral", "partly agree" and "strongly agree". A text field was also included for feedback, where respondents could, if they wished, describe their opinions on the prototype more clearly. It should also be noted that the survey was designed to be answered anonymously. Table 1 presents the statements included in the survey.

VI. RESULTS AND DISCUSSION

The survey was sent to 63 partners in total, 3 of whom responded. Although the survey was answered by around 5% of the total number of people to whom it was sent, it should be reiterated that part of the validation of the prototype was carried out in meetings with the partners throughout the development of the project.

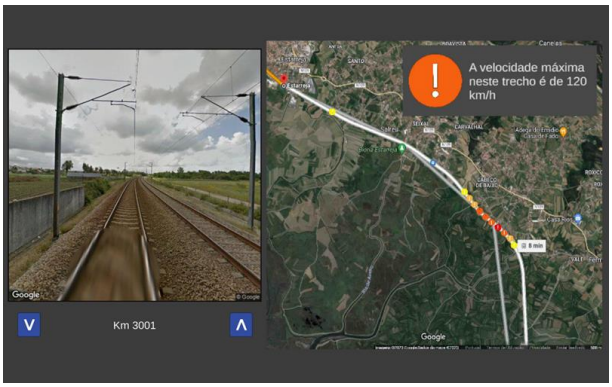

Figure 5. Railroad infrastructure's mobile dashboard

TABLE I. EVALUATION SURVEY STATEMENTS

#	Statement
S1	The user interface of the prototype is, in general, intuitive and easy to interact with.
S2	The data on the damage indicators is presented in a clear and understandable way.
S3	The data shown by the history graph of the damage indicators is presented in a clear and comprehensible manner.
S4	If employed in a real-world context, the proposed prototype would be useful for supporting the monitoring of the conditions of rail transport vehicles.
S5	If employed in a real-world context, the proposed prototype would be useful for supporting preventive maintenance.

The participants' answers to items S1 to S3 indicate a generally positive opinion about the user interface of the mobile application, although some issues were raised regarding the clarity of the information and navigation, which still need to be improved. The feedback given by one of the respondents mentions that, when the application launches, the user interface displays placeholder texts in the dropdown menus and in the sections where the names of the component and the chosen variable will be displayed. These texts consist of generic words such as "Train", "Component Name" and "Variable". This could make the interface less intuitive, as it creates confusion among users. Changing the placeholder text to explanatory phrases - such as "Select variable" instead of "Variable" - could help make the user interface more intuitive.

TABLE II. EVALUATION SURVEY RESULTS

#	Respondent 1	Respondent 2	Respondent 3
S1	5	4	5
S2	5	4	4
S3	5	4	4
S4	4	5	5
S5	4	4	5

 ISBN: 978-1-68558-096-4

A different participant also mentioned that they would like to be able to "click" on the vehicle component they want to analyze. Although the meaning of the term "click" was not entirely clear in this case, it is understood that the respondent would like to select the vehicle component directly on the CAD model of the vehicle. This functionality already exists and it is essential for the interaction with the application, as it is the only way to select components. This means that the need for direct interaction with the CAD model of the train is not obvious to some users. The inclusion of an explanatory text - such as "Touch the train to select the component you want to analyze" - could eliminate this issue.

Despite receiving a positive evaluation from the respondents, the history graph of the damage indicators also has some points that need to be improved. It was mentioned that the way the graph is presented would not be appropriate, as continuous lines are used to illustrate variations in measurements. These lines, according to feedback from one of the respondents, do not represent real variations and, because of this, broken lines should be used instead. In fact, the value variations illustrated in the graph do not correspond to reality, which could lead to confusion among users and even result in misunderstandings in decision-making if the prototype was employed in a real-world scenario. As stated by the respondent, the use of broken lines would be the most appropriate method.

The participants also mentioned that it would be useful to include a button for viewing measurements prior to those displayed on the graph. The functionality for navigating along the graph is already implemented and is done by tapping the right and left corners of the graph's window. However, due to the lack of buttons or visual indications alerting users to the existence of this functionality, it can go unnoticed. This shortcoming could be remedied by simply adding buttons with arrows pointing to the right and left, positioned in the right and left corners of the graph, respectively.

Finally, the opinions expressed by the respondents in items 4 and 5 indicate that the proposed digital twin prototype shows potential for supporting the monitoring of railway vehicles and the implementation of preventive maintenance in a real-world context. The feedback indicates that, although some corrections are needed in relation to the user interface of the mobile application, the prototype displays the essential characteristics of a digital twin.

VII. Conclusion and Future Work

The present work was designed to explore the potential of the digital twin concept in supporting rail transport operations, particularly with regard to monitoring and preventive maintenance. This objective was achieved by developing digital twin prototypes of a railway vehicle and of a section of railroad, which provided a greater understanding of how the digital twin concept can offer a more complete and functional view of the operational conditions of railway equipment during operation. It was also possible to see how this concept supports the implementation of preventive maintenance processes for vehicles, by making it possible to visualize the evolution of damage indicators relating to the vehicle's components.

Although the proposed prototypes allow us to see the potential of the digital twin concept in the context of rail transport, there are several improvements that could be implemented. Among them is the incorporation of functionalities for sending and displaying warnings, through which users would be able to alert others to potential defects or physical irregularities in the vehicles and in the railroad infrastructure. The mobile application would allow users to submit alert notes, which would be stored in a database and associated to the corresponding equipment.

Another improvement that could be implemented is the real-time collection and display of vehicle location data using Global Positioning Systems (GPS) sensors. With this method, it would be possible to identify the exact location of a vehicle along its route, as well as allow the subsequent display of information about the schedule of the vehicles, such as the time taken to complete a given route.

The present work could also encourage interest in exploring the potential of other technologies in the Industry 4.0 paradigm - such as augmented reality and algorithms for analyzing big data and computer vision - in the context of rail transport services. Some of these technologies could be used, for example, to support the maintenance processes of train components: by using a mobile application, operators would be able to view interactive guides - generated by computer vision algorithms - and maintenance instructions in augmented reality.

Acknowledgment

This work was developed as part of the Ferrovia 4.0 research project, which was co-financed by COMPETE 2020, Portugal 2020, Lisboa 2020 and the European Union's European Structural and Investment Funds.

References

[1] Y. Liao, F. Deschamps, E. d. F. R. Loures, and L. F. P. Ramos, "Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal," International Journal of Production Research, vol. 55, no. 12, pp. 3609-3629, 06/18 2017, doi: 10.1080/00207543.2017.1308576.

[2] H. Kagermann, W. Wahlster, and J. Helbig, "Recommendations for implementing the strategic initiative INDUSTRIE 4.0," 2013. [Online]. Available: https://en.acatech.de/publication/recommendations-for-implementing-the-strategic-initiative-industrie-4-0-final-report-of-the-industrie-4-0-working-group/.

[3] F. Shrouf, J. Ordieres, and G. Miragliotta, "Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm," in 2014 IEEE International Conference on Industrial Engineering and Engineering Management, 9-12 Dec. 2014, pp. 697-701, doi: 10.1109/IEEM.2014.7058728.

[4] Q. Qi et al., "Enabling technologies and tools for digital twin," Journal of Manufacturing Systems, vol. 58, pp. 3-21, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.jmsy.2019.10.001.

[5] K. Y. H. Lim, P. Zheng, and C.-H. Chen, "A state-of-the-art survey of Digital Twin: techniques, engineering product lifecycle management and business innovation perspectives,"

 ISBN: 978-1-68558-096-4

Journal of Intelligent Manufacturing, vol. 31, no. 6, pp. 1313-1337, 2020/08/01 2020, doi: 10.1007/s10845-019-01512-w.

[6] S. Jeschke and R. Grassmann, "Development of a Generic Implementation Strategy of Digital Twins in Logistics Systems under Consideration of the German Rail Transport," Applied Sciences, vol. 11, no. 21, doi: 10.3390/app112110289.

[7] Projeto Ferrovia 4.0, "Anexo Técnico - Ferrovia 4.0," 2019.

[8] M. Grieves, "Digital Twin: Manufacturing Excellence through Virtual Factory Replication," 03/01 2015.

[9] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art," IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 2405-2415, 2019, doi: 10.1109/TII.2018.2873186.

[10] E. J. Tuegel, A. R. Ingraffea, T. G. Eason, and S. M. Spottswood, "Reengineering Aircraft Structural Life Prediction Using a Digital Twin," International Journal of Aerospace Engineering, vol. 2011, p. 154798, 2011/10/23 2011, doi: 10.1155/2011/154798.

[11] E. H. Glaessgen and D. S. Stargel, "The digital twin paradigm for future NASA and U.S. air force vehicles," in 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 2012, 2012. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84881388851&partnerID=40&md5=76921d9a4627f52dfccb21e0f7a9d767. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84881388851&partnerID=40&md5=76921d9a4627f52dfccb21e0f7a9d767

[12] J. Guo, N. Zhao, L. Sun, and S. Zhang, "Modular based flexible digital twin for factory design," Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 3, pp. 1189-1200, 2019/03/01 2019, doi: 10.1007/s12652-018-0953-6.

[13] J. Vachalek, L. Bartalsky, O. Rovny, D. Sismisova, M. Morhac, and M. Loksik, "The digital twin of an industrial production line within the industry 4.0 concept," in Proceedings of the 2017 21st International Conference on Process Control, PC 2017, 2017, pp. 258-262, doi: 10.1109/PC.2017.7976223. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027512911&doi=10.1109%2fPC.2017.7976223&partnerID=40&md5=31bb2e758a775ddea30b1d66b905d3b7

[14] F. Tao and M. Zhang, "Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing," IEEE Access, Article vol. 5, pp. 20418-20427, 2017, Art no. 8049520, doi: 10.1109/ACCESS.2017.2756069.

[15] S. Haag and R. Anderl, "Digital twin – Proof of concept," Manufacturing Letters, Article vol. 15, pp. 64-66, 2018, doi: 10.1016/j.mfglet.2018.02.006.

[16] J. Liu, H. Zhou, G. Tian, X. Liu, and X. Jing, "Digital twin-based process reuse and evaluation approach for smart process planning," International Journal of Advanced Manufacturing Technology, Article vol. 100, no. 5-8, pp. 1619-1634, 2019, doi: 10.1007/s00170-018-2748-5.

[17] The AnyLogic Company. "Alstom Develops a Rail Network Digital Twin for Railway Yard Design and Predictive Fleet Maintenance." https://www.anylogic.com/resources/case-studies/digital-twin-of-rail-network-for-train-fleet-maintenance-decision-support/ (accessed September 4, 2023).

[18] Unity Technologies. "Unity Real-Time Development Platform | 3D, 2D VR & AR Engine." https://unity.com/ (accessed September 4, 2023).

[19] R. Göb, C. McCollin, and M. F. Ramalhoto, "Ordinal methodology in the analysis of likert scales," Quality and Quantity, Article vol. 41, no. 5, pp. 601-626, 2007, doi: 10.1007/s11135-007-9089-z.

[20] F. Paz and J. A. Pow-Sang, "A systematic mapping review of usability evaluation methods for software development process," International Journal of Software Engineering and its Applications, Article vol. 10, no. 1, pp. 165-178, 2016, doi: 10.14257/ijseia.2016.10.1.16.

 ISBN: 978-1-68558-096-4