David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

7th International Conference on Advances in Computer Science and Information Technology (ACSTY 2021),
March 20-21, 2021, Vienna, Austria.

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-37-4
DOI: 10.5121/csit.2021.110301- 10.5121/csit.2021.110309

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 7th International Conference on Advances in Computer Science and Information Technology (ACSTY 2021), March 20-21, 2021, Vienna, Austria, 2nd International Conference on Artificial Intelligence and Big Data (AIBD 2021), 2nd International Conference on Machine Learning and Soft Computing (MLSC 2021), 2nd International Conference on Cloud Computing and IOT (CCCIOT 2021), 7th International Conference on Natural Language Processing (NATP 2021) and 7th International Conference on Software Engineering (SOFE 2021) was collocated with 7th International Conference on Advances in Computer Science and Information Technology (ACSTY 2021). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ACSTY 2021, AIBD 2021, MLSC 2021, CCCIOT 2021, NATP 2021 and SOFE 2021 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, 2021, AIBD 2021, MLSC 2021, CCCIOT 2021, NATP 2021 and SOFE 2021 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the 2021, AIBD 2021, MLSC 2021, CCCIOT 2021, NATP 2021 and SOFE 2021.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students andeducators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years tocome.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair

Organization

David C. Wyld, Southeastern Louisiana University, USA
DhinaharanNagamalai(Eds) Wireilla Net Solutions, Australia

Program Committee Members

Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Abderrahmane EZ ZAHOUT, Mohamed 5 University, Morocco
Abdullah, Adigrat University, Ethiopia
Abhishek Appaji, B.M.S. College of Engineering, India
Abhishek Shukla, A P J Abdul Kalam Technical University, India
Addisson Salazar, Universitat Politècnica de València, Spain
Afaq Ahmad, Sultan Qaboos University, Oman
Ahmed Elngar, Beni-Suef University, Egypt
Ahmed Kadhim Hussein, Babylon University, Iraq
AKhil Gupta, Lovely Professional University, India
Ali Asghar Rahmani Hosseinabadi, University of Regina, Canada
Alireza Valipour Baboli, Technical and Vocational University, Iran
Amina El murabet, Abdelmalek Essaadi University, Morocco
Anirban Banik, National Institute of Technology Agartala, India
Ann Zeki Ablahd, Northern Technical University, Iraq
Antonio Carlos Bento, Anhembi Morumbi University, Brazil
Antonio Moreira, University of Aveiro, Portugal
Apu Kumar Saha, National Institute of Technology Agartala, India
Aref Wazwaz, Dhofar University, Oman
Asif Irshad Khan, Kind Abdulaziz University, Saudi Arabia
Attila Kertesz, University of Szeged, Hungary
Ayad Salhieh, Australian College of Kuwait (ACK), Kuwait
Azeddine Wahbi, Hassan II University, Morocco
Bahaa Saleh, ICT, Egypt
Balagadde, Kampala International University, Uganda
Bilal Alatas, Firat University, Turkey
Bin Zhao, Xidian University, China
Boukarinassim, Skikda University, Algeria
Brijender Kahanwal, PT. C.L.S. Government College, India
Chandra, San Jose State University, USA
Chang-Yong Lee, Kongju National University, South Korea
Charalampos Karagiannidis, University of Thessaly, Greece
Ching-Nung Yang, National Dong Hwa Univesity, Taiwan
Christian Mancas, Ovidius University, Constanta, Romania
Christos Bouras, University of Patras, Greece
Claudia Canali, University of Modena and Reggio Emilia, Italy
CS Chin, Newcastle University, Singapore
Dadmehr Rahbari, University of Qom, Iran
Daniel Ekpenyong Asuquo, University of Uyo, Nigeria
Daniel Hunyadi, "Lucian Blaga" University of Sibiu, Romania

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

7th International Conference on Advances in Computer Science and Information Technology (ACSTY 2021)

2nd International Conference on Artificial Intelligence and Big Data (AIBD 2021)

2nd International Conference on Machine Learning and Soft Computing (MLSC 2021)

2nd International Conference on Cloud Computing and IOT (CCCIOT 2021)

7th International Conference on Natural Language Processing (NATP 2021)

7th International Conference on Software Engineering (SOFE 2021)

FINDING SIMILAR ENTITIES ACROSS KNOWLEDGE GRAPHS

Sareh Aghaei and Anna Fensel

Semantic Technology Institute (STI) Innsbruck, Department of Computer Science, University of Innsbruck, Innsbruck, Austria

ABSTRACT

Finding similar entities among knowledge graphs is an essential research problem for knowledge integration and knowledge graph connection. This paper aims at finding semantically similar entities between two knowledge graphs. This can help end users and search agents more effectively and easily access pertinent information across knowledge graphs. Given a query entity in one knowledge graph, the proposed approach tries to find the most similar entity in another knowledge graph. The main idea is to leverage graph embedding, clustering, regression and sentence embedding. In this approach, RDF2Vec has been employed to generate vector representations of all entities of the second knowledge graph and then the vectors have been clustered based on cosine similarity using K medoids algorithm. Then, an artificial neural network with multilayer perception topology has been used as a regression model to predict the corresponding vector in the second knowledge graph for a given vector from the first knowledge graph. After determining the cluster of the predicated vector, the entities of the detected cluster are ranked through sentence-BERT method and finally the entity with the highest rank is chosen as the most similar one. To evaluate the proposed approach, experiments have been conducted on real-world knowledge graphs. The experimental results demonstrate the effectiveness of the proposed approach.

KEYWORDS

Knowledge Graph, Similar Entity, Graph Embedding, Clustering, Regression, Sentence Embedding.

1. INTRODUCTION

With the rise of knowledge graphs (KGs), interlinking KGs has attracted a lot of attention. A KG is a huge semantic net which integrates various, inconsistent and heterogeneous information resources to represent knowledge about different domains [1]. KGs have proven beneficial for artificial intelligence applications, including question answering, document retrieval, recommendation systems and knowledge reasoning [2, 3]. To interlink KGs, it is crucial to find similar entities across the KGs that have high semantic similarity to each other [3]. Addressing this challenge would allow end users and search agents to find more relevant information across KGs [3]. This can be used in different applications, such as online marketing, search engine optimisation and online services provisioning, for example, in tourism [4].

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 01-11, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110301

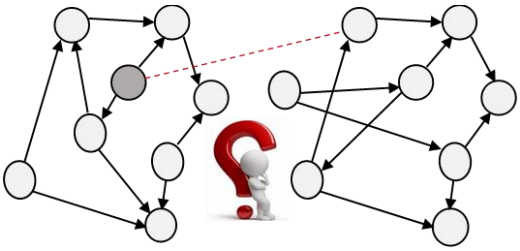


Figure 1. The interlinking problem over the knowledge graphs

This paper delves into the problem of finding similar entities across different KGs. Given a query entity in one KG, this study aims to find the most similar entity in another KG as illustrated in Figure 1. Here, the entity pair may not reference the same real-world entity but have the most similarity to each other. The proposed approach includes four main steps: graph embedding, clustering, regression, sentence embedding as showed in Figure 2.

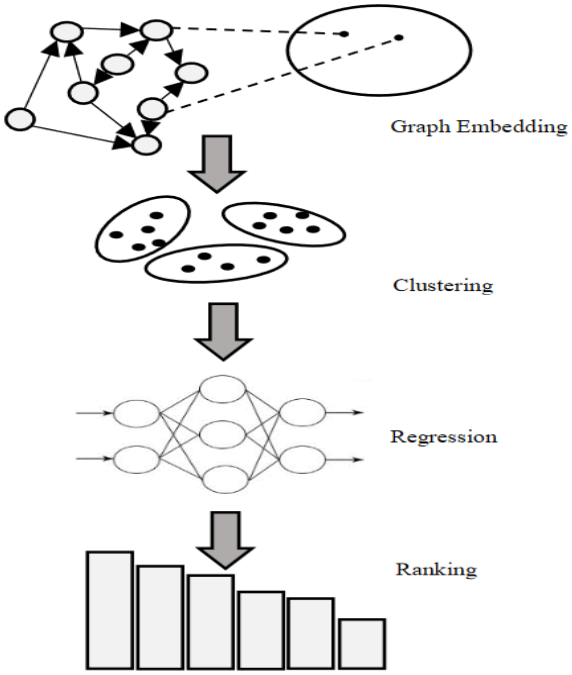


Figure 2. The proposed approach

A graph embedding is used to represent entities of a KG in low dimensional semantic space while preserving the structural as well as the semantic features of the entities. Recently, different graph embedding techniques have been proposed to capture different aspects of graphs. In this paper, RDF2VEC graph embedding technique has been applied to capture the semantic similarity of entities in each RDF KG. RDF2VEC adapts the language modelling approach of word2vec to RDF graph embeddings [5]. RDF2Vec converts the KG to a set of sequences (using graph walks and Weisfeiler-Lehman subtree RDF graph kernels) and then trains a neural network model to learn vector representation of entities. It maps each entity to a low dimensional vector of latent numerical values in which semantically and syntactically closer entities will appear closer in the vector space [5, 6].

Clustering for interlinking large-scale KGs is a fundamental step. Although there are different approaches for clustering large amounts of data, the proposed approach uses K means and K medoids clustering based on two different metrics, Euclidean distance and cosine distance, to group vector representations of the second KG. In these algorithms, vectors are segmented into different groups where each cluster contains at least one vector. No vectors may be placed into more than one cluster. Furthermore, the number of clusters 'K' must be specified prior to initiating the algorithm and also, they allow for interpretability of the cluster centres. K-means targets to minimize the total squared error from a central position in each cluster namely centroid. Whereas K medoids aims to minimize the sum of dissimilarities between vectors labelled to be in a cluster and one of the vectors considered as the representative of that cluster called medoid [7, 8].

Various methods, including a variety of regression techniques and artificial neural networks, can be applied to develop a forecasting model. The present approach has employed artificial neural network and multivariate multiple linear regression techniques to predict the vector representation for a given embedding from the first KG. The neural network technique has become an increasingly popular modelling tool for forecasting. Multilayer perceptron (MLP) with back-propagation learning rule is adopted to predict the embeddings of the second KG according to the embeddings of the first KG. Furthermore, the multivariate multiple linear regression model is beneficial in discovering the association between various independent and dependent variables. It attempts to model the correlation between involving variables and response variables depending on linear equation into the observed data [9].

The entity description and other textual values of properties in KG usually carry conceptual semantic information [10]. Based on the entity description, the Sentence-BERT technique is adopted to compute the textual similarity. Sentence-BERT (S-BERT) is a modification of the pretrained BERT network that employs siamese and triplet networks in order to derive semantically meaningful sentence embeddings [11]. The derived sentence embeddings of the entities of the chosen cluster are compared with the sentence embedding of the given entity and ranked based on cosine-similarity. Finally, the entity ranked first is selected.

The approach of this paper can take advantage of value-oriented and record-oriented [12] techniques. According to [12], value-oriented techniques compute the similarity between entities on the attribute level and record-oriented techniques contain solutions based on learning, rules, contexts. Furthermore, it works independent of mapping schema and benefits the structure of KGs.

The remainder of this paper is structured as follows. The next section presents some related studies. In section 3, the proposed approach is presented. Section 4 demonstrates the results obtained and evaluation. Finally, concluding remarks and an outlook on future work are in Section 5.

2. RELATED WORKS

The task of interlinking KGs aims to find entities in two KGs that have semantic relations. The different KGs are constructed independently from each other, so they contain complementary entities. While numerous studies exist regarding entity alignment (also named entity resolution, duplicate detection, record linkage, or entity resolution) with the goal of finding entities from different KGs that refer to the same real-world identity [9], there is a lack of approaches to find entities with the most similarity so that those entities may not be the same entity pairs.

SILK [13], LIMES [14] and Dude [15]are examples of traditional approaches which have leveraged different similarity metrics including string similarity, numeric similarity, date similarity, word relation and fuzzy string similarity. These approaches usually have an ability to build more complex similarity metrics through combining the similarity metrics for increasing their functionality and performance.

In [3], a classification-based approach has been provided to address the entity alignment problem between source and target KGs. Using source/target entity pairs, a classifier is trained and the probability of predicting an alignment is adopted for candidate ranking. RDF2Vec graph embedding technique has been used to the embeddings of the source and target entities, then the embedding of the given entity in the source KG and the candidate entity in the target KG are concatenated into one feature vector and fed into a multi-layer perception. Finally, it sorts the candidates by the match probability for evaluation.

MtransE [16] which is a multi-lingual KG embedding model has consisted of two component models, called knowledge model and alignment model, to learn the multilingual KG structure. The knowledge model encodes entities by adopting TransE [17]. On top of that, the alignment model employs three different techniques to learn cross-lingual alignment for entities and relations, namely distance-based axis calibration, translation vectors, and linear transformations. Comparisons across the used techniques show that the linear-transformation-technique based on different loss functions.

A KG alignment network, namely AliNet [18] has been proposed to reduce the non-isomorphism of neighbourhood structures in an end-to-end manner. Since the schema heterogeneity ensures dissimilarity across counterpart entities, AliNet introduces distant neighbours to expand the overlap between their neighbourhood structures using an attention mechanism. The neighbourhood information within multiple hops are captured through the applied gating mechanism in each layer.

For cross-lingual entity alignment, a joint attribute-preserving embedding model has been introduced to jointly embed the structures of two knowledge bases into a unified vector space and then refine it through leveraging attribute correlations in the knowledge bases. This model has utilized the structure embedding and attribute embedding in order to represent the relationship structures and attribute correlations of knowledge bases and learn approximate embeddings for latent aligned entities [19].

REA [20] has proposed a framework for robust entity alignment over KGs. The framework consists of two components: noise detection and noise-aware entity alignment. In order to encode the information of KGs, it leverages a graph neural network-based encoder. The noise-aware entity alignment component targets to diminish the distance between two entities in a labelled entity pair to avoid the noise based on the encoder. The idea of the noise detection component is to generate noisy data and have an ability to differentiate between the generated noisy data and real data following the adversarial training principle. However, REA cannot distinguish a few real entity pairs with real pairs in some cases.

3. APPROACH

Problem Definition – A Resource Description Framework(RDF)KG can be denoted as $G = (E, R, T)$, where E is the set of entities, R is the set of relations, and T is the set of triples. A KG triple $(e_h, r, e_t$ indicates the head entity e_h is linked to the tail entity e_t by the relation r. Let $G_1 = (E_1, R_1, T_1)$ and $G_2 = (E_2, R_2, T_2)$ be the first and second KG, respectively. The task is to find the

entity $e_2 \in E_2$ which has the most semantic similarity to the given entity $e_1 \in E_1$ from the first KG, thus $\forall e_1 \in E_1 : \exists e_2 \in E_2 that\ e_2 \approx e_1$.

Methodology - The proposed approach includes four main steps: graph embedding, clustering, regression, sentence embedding. In the first step, RDF2Vec [6] algorithm has been used to generate RDF graph embeddings. The generated vector representations of the second KG are clustered in the next step. Then, a regression model is trained according to the vector representations of the same entities between the first and second KGs. For each given entity of the first KG, the correspondent vector from the second KG is predicated and its cluster is determined. In the final step, the sentence embedding is utilized based on the value of description property in the predicated cluster by BERT and the generated vectors are ranked based on cosine-similarity with the sentence vector of the source entity. The target entity with top rank is the entity with more similarity.

Below, the approach steps including graph embedding, clustering, regression and rankling are described in detail.

3.1. Graph Embedding

RDF2Vec, which is a technique to embed RDF graphs for learning latent numerical representations of entities in RDF graphs, has been inspired by the word2vec approach. The Word2vec is a particularly computationally-efficient two-layer neural language model to generate word embeddings from raw text [6, 21]. The Word2vec takes a set of sentences as input, and trains a two-layer neural network using one of the two algorithms, the continuous bag of words model (CBOW) and the skip-gram model (SG). The CBOW predicts a target word from its context within a given window and the SG predicts the context words given a word. The RDF2Vec first converts the RDF graphs in a set of sequences using two techniques, Weisfeiler-Lehman Subtree RDF Graph Kernels and graph walks, which are then used as input for the word2vec algorithm to train the neural language model [21]. When the training is done, all entities are projected into a lower-dimensional feature space, and semantically similar entities are closer in the vector space than dissimilar ones. For more details the readers are referred to [6, 21]. In the proposed approach, RDF2Vec is used to generate embeddings for all entities of the second KG, the entity pairs which have the same relation between the first and second KGs and each given entity from the first KG.

3.2. Clustering

Clustering is of key importance for interlinking entities from multiple KG. To achieve high efficiency for large KGs, interlinking solutions have to avoid comparing each entity to all other entities. This can be gained by so-called blocking strategies where only entities within the same cluster (block) need to be compared with each other [22]. Clustering algorithms typically try to cluster entities such that the similarity between entities within a cluster is maximized while the similarity between entities of different clusters is minimized [22].

In proposed approach, the K medoids algorithm has been adopted to cluster vector representations of the second KG. This algorithm is relatively simple to implement and scales to large KGs. Moreover, the medoids of the K clusters can be used to determine the relevant cluster of new vectors. The cosine similarity has been chosen to group together all close vectors of the second KG.

3.3. Regression

The objective of the regression prediction model is to find the transitions between the vector spaces of the first and second KGs [16]. Since the embeddings of the KGs are learned separately, it is essential to learn correspondences between two semantic spaces. One feasible solution to the dilemma is to estimate regression relationships between the entities of the first KG and the entities of the second KG based on existing similar entities.

In this step, the following regression prediction models have been applied:

Multi-layer perceptron (MLP) network - One of the most popular artificial neural networks which can be used to find associations between two sets of variables, is the feed-forward multi-layer network, which uses a back-propagation learning algorithm. It consists of one or more hidden layer(s), containing computational nodes named neurons/perceptrons which intervene between input and output of the network, and can improve the accuracy of the network [23].

Multivariate multiple linear regression (MMLR) - The multivariate multiple linear regression is a statistical method that allows to predict of several dependent variables from a set of independent variables and its purpose is finding the best fitting line which is called regression function [24, 25].

In the proposed approach, a MLP network and also a MMLR is used to predict the embedding of similar entity in the second KG based on the embedding of the given entity of the first KG. The entity pairs which have same relation between the first and second KGs have been considered as training data to train the regression models.

3.4. Ranking

Sentence-BERT (SBERT) can be considered a modification of the pretrained BERT network which generates a fixed sized sentence embedding by adding a pooling operation to the output of BERT / RoBERTa. In order to fine-tune BERT / RoBERTa, SBERT uses siamese and triplet networks to update the weights such that the generated sentence embeddings are semantically meaningful and can be compared with cosine-similarity [11].

In the proposed approach, the textual values of entity properties (e.g. description) are adopted to input sentences for SBERT. The sentence embeddings of the determined cluster in the previous step are compared with the sentence embedding of the given entity from the first KG and then ranked based on cosine similarity. Finally, the highest ranked entity is chosen as the entity which has the most semantic similarity with the given one.

4. EVALUATION

In order to evaluate the approach presented in this paper, DBPedia [26] and SalzburgerLand [27] KGs have been used as the first and second KGs. The SalzburgerLand KG is a KG describing touristic entities of the region of Salzburg, Austria, and among others it includes 21496 triples and 571 entities which reference DBPedia KG, which is a KG representing Wikipedia. The evaluation code has been written in Python and is publicly available athttps://github.com/sareaghaei/interlinking.

For RDF2Vec graph embedding, the depth of graph walks and the limit number of walks per entity are 8 and 20, respectively. The outcome of this step is a 100-dimensional vector for each entity.

K means and K medoids algorithms have been employed to group together all close vectors of SalzburgerLand KG based on the Euclidean distance and the standard cosine similarity, respectively. In practice, for obtaining the best clustering quality, the optimal value of K is determined by experiments (K = 2). Not only does K medoids clustering has higher score in terms of silhouette coefficient, but also leads to better result in the next steps. The figure for silhouette coefficient is 0.60 and 0.39 in K medoids and K means, respectively. Figure 3 illustrates the sets of clustering. Also, the centroids and medoids have been shown in green colour and bigger size.

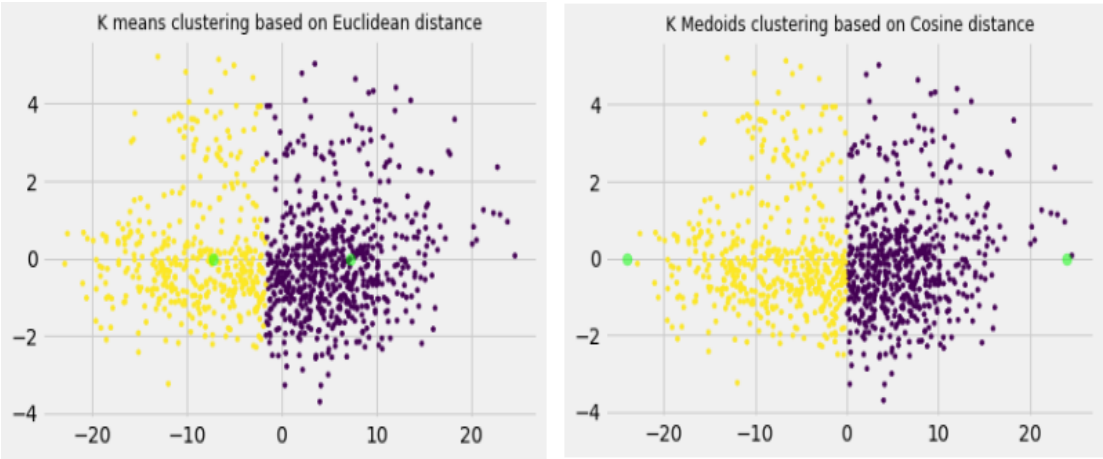


Figure 3. The clusters of K means and K medoids algorithms

Note that principal component analysis (PCA) [28] has been used to automatically perform dimensionality reduction over the embeddings before visualizing the clusters in Figure 3. The Scikit-Learn library, which has implemented the PCA technique, applies the full singular value decomposition (SVD) or a randomized truncated SVD depending on the shape of the input data and the number of components to extract [29]. Here, PCA has been used to reduce dimensions from 100 to 2.

A multi-layer perception (MLP) with 1 hidden layer which has size 50 using the ReLU activation function, followed by a fully-connected layer and ReLU to output the final prediction has been applied as the regression predication model. The model is trained using the Adam optimizer. Moreover, a multivariate multiple linear regression model has been trained in which the DBPedia KG and the SalzburgerLand KG embeddings are considered as independent and dependent variables, respectively. In order to provide a more complete and effective evaluation of the regression models, the cross-validation has been performed using K-fold algorithm with 5-folds.

The smaller the difference between the predicted vectors and the real vectors, the higher the prediction accuracy that the models provide. Thus, mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) have been applied to measure the performance of the models. MAE represents the average of the absolute difference between the actual and predicted values, MSE is defined as average of the square of the difference between actual and predicted values and RMSE is the square root of mean squared error which computes the

standard deviation of residuals. Mathematical formulas to calculate these metrics can be written as following where $y^{\wedge}$ is the predicated value of y:

$$MAE = \frac{1}{n}\sum_{i=1}^{i=n} |y_i - y_i^{\wedge}|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{i=n} (y_i - y_i^{\wedge})^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n} (y_i - y_i^{\wedge})^2}$$

Overall, the MLP network outperforms the MMLR model for predicting the embeddings of SalzburgerLand KG, the figures for MAE and MSE in the MLP network are 0.866 and 1.005, respectively, whereas those of the MMLR model are 0.966 and 1.348, the evaluation of results is shown in Figure 4.

Sentence-Transformers which is a Python framework has been used to compute sentence / text embeddings [11]. It is based on PyTorch and Transformers and the produced sentence embeddings are 150-dimensional vectors which have been compared with cosine-similarity in order to be ranked.

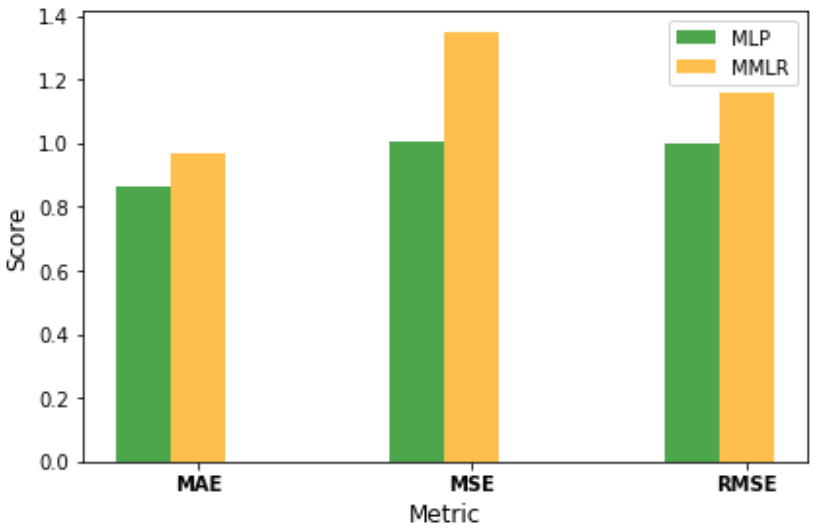


Figure 4. The evaluation of prediction errors

5. CONCLUSION

This paper proposes an approach to interlink KGs. In order to find the most similar entity from a KG (second KG) with a given entity from another KG (first KG), the proposed approach includes four steps: graph embedding, clustering, regression and ranking. RDF2Vec technique is used to generate vector representations and then K means/K medoids algorithms are adopted for clustering of the embeddings of the second KG. To learn associations between distinct semantic spaces (one from each KG), multi-layer perceptron networks and multivariate multiple linear regressions are trained and used to predict the embedding from the second KG based on the embedding of the given entity from the first one. By comparing the predicted vector with the centroid-medoid of the clusters, the correspondent cluster is determined and its entities are ranked based on cosine similarity between their sentence embedding and the sentence embedding of the given entity. SBERT is used to compute sentence embeddings of the entities over their textual values of the properties. The experimental results show that the proposed approach as one of the state-of-art interlinking approaches can achieve high accuracy. However, in the proposed approach, the regression model requires training based on the entity pairs between two KGs, it can definitely be considered a drawback due to lack of the pairs in some cases. For future work, aside from experimenting with other embedding learning techniques for KGs, learning associations on KGs with better accuracy and experiments on different KGs are planned.

ACKNOWLEDGEMENTS

This work has been partially funded by the project WordLiftNG within the Eureka, Eurostars Programme (grant agreement number 877857 with the Austrian Research Promotion Agency (FFG)).

REFERENCES

[1] Dieter Fensel, Umutcan Simsek, Kevin Angele, Elwin Huaman, Elias Kärle,Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler(2020) “Introduction: What Is a Knowledge Graph?”, Springer International Publishing, pp. 1–10.

[2] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu (2018) “Bootstrapping entity alignment with knowledge graph embedding”, International Joint Conferences on Artificial Intelligence, pp. 4396-4402.

[3] Michael Azmy, Peng Shi, Jimmy Lin, and Ihab F. Ilyas (2019) “Matching entities across different knowledge graphs with graph embeddings”, CoRR, abs/1903.06607.

[4] Anna Fensel, Zaenal Akbar, Elias Kärle, Christoph Blank, Patrick Pixner, and Andreas Gruber (2020) “Knowledge Graphs for Online Marketing and Sales of Touristic Services”, Information, 11(5), 253.

[5] RemziCelebi, HuseyinUyar, Erkan Yasar, OzgurGumus, OguzDikenelli, and Michel Dumontier (2019) “Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings”, BMC Bioinformatics.

[6] PetarRistoski, and Heiko Paulheim (2016) “RDF2Vec: RDF Graph Embeddings for Data Mining.” International Semantic Web Conference.

[7] Tagaram Soni Madhulatha (2011) “Comparison between K-Means and K-Medoids Clustering Algorithms”, Advances in Computing and Information Technology, pp. 472-481.

[8] Preeti Arora, DeepaliVirmani, and Shipra Varshney (2016) “Analysis of K-Means and K-Medoids Algorithm For Big Data”, Procedia Computer Science, Vol. 78, pp. 507-512.

[9] Elwin Huaman, Elias Kärle, and Dieter Fensel (2020) “Duplication Detection in Knowledge Graphs: Literature and Tools”, arXiv:2004.08257.

[10] Ying Shen, Kaiqi Yuan, Jingchao Dai, Buzhou Tang, Min Yang, and Kai Lei (2019) “KGDDS: A System for Drug-Drug Similarity Measure in Therapeutic Substitution based on Knowledge Graph Curation”, Journal of medical systems 43, 92.

[11] Nils Reimers, and Iryna Gurevych (2019) "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982-3992.

[12] Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Gaia Varese (2011) "Ontology and instance matching. In Knowledge-Driven Multimedia Information Extraction and Ontology Evolution", Springer, pp. 167-195.

[13] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov (2009) "Discovering and Maintaining Links on the Web of Data", In Proceedings of The International Semantic Web Conference (ISWC) ISWC, pp. 650-665.

[14] Axel-CyrilleNgongaNgomo, and Sören Auer (2011) "LIMES - A time-efficient approach for large-scale link discovery on the web of data", In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI2011), pp. 2312-2317.

[15] Lars Marius Garshol, and Axel Borge (2013) "Hafslundsesam - an archive on semantics", In Proceedings of the 10th Extending Semantic Web Conference (ESWC2013), vol. 7882, pp. 578-592.

[16] Muhao Chen, Yingtao Tian, MohanYang, and Carlo Zaniolo (2016) "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment", arXivpreprint arXiv:1611.03954.

[17] Antoine Bordes,NicolasUsunier, Alberto Garcia-Duran, Jason Weston, and OksanaYakhnenko (2013) "Translating embeddings for modelling multi-relational data", In Proceedings of the 26th International Conference on Neural Information Processing Systems, pp. 2787-2795.

[18] Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu (2019) "Knowledge graph alignment network with gated multi-hop neighborhood aggregation", arXiv:1911.08936.

[19] Zequn Sun, Wei Hu, and Chengkai Li (2017) "Cross-lingual entity alignment via joint attribute-preserving embedding", In Proceedings of The International Semantic Web Conference (ISWC) ISWC, vol. 10587, pp. 628-644.

[20] Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang (2020) "Rea: Robust cross-lingual entity alignment between knowledge graphs", In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery& Data Mining, pp. 2175-2184.

[21] PetarRistoski, Jessica Rosati, Tommaso Di Noia, Renato De Leone, and Heiko Paulheim (2018) "RDF2Vec: RDF Graph Embeddings and Their Applications", Semantic Web, Vol. 10, No. 4, pp. 721-752.

[22] Ali Saeedi, Markus Nentwig, Eric Peukert, and Erhard Rahm (2018) "Scalable matching and clusteringof entities with famer", Complex Systems Informatics and Modelling Quarterly, pp. 61-83.

[23] M.E. Hamzehie, S Mazinani, F. Davardoost, A. Mokhtare, H. Najibi, BVdBruggen, and S. Darvishmanesh (2014) "Developing a feed forward multilayer neural network model for prediction of CO2 solubility in blended aqueous amine solutions", Journal of Natural Gas Science and Engineering, pp. 19-25.

[24] Lianpeng Li, Jian Dong, Decheng Zuo, and Jin Wu (2019) "SLA-aware and energy-efficient VM consolidation in cloud data centers using robust linear regression prediction model", IEEE Access 7, pp. 9490-9500.

[25] Yanming Li, Bin Nan, and Ji Zhu (2015) "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure", Biometrics, 71, pp. 354-363.

[26] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer (2013) "A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", Semantic Web Journal, pp. 167-195.

[27] "Welcome to SalzburgerLand Data Hub", Accessed on: Oct. 18, 2020. [Online]. Available: http://data.salzburgerland.com/dataset/salzburgerland-en.

[28] IanT. Jolliffe, and Jorge Cadima (2002) "Principal Component Analysis", Wiley Online Library.

[29] "Principal component analysis (PCA)", Accessed on: Oct. 5, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/decomposition.html#pca.

AUTHORS

Sareh Aghaei received the master's degree in computer engineering from the University of Isfahan, Iran and is currently a PhD student at the University of Innsbruck, Austria. Her research areas include semantic web, knowledge graphs and question answering systems.

Anna Fensel is Associate Professor at the University of Innsbruck, Austria. Earlier she worked as a Senior Researcher at FTW – Telecommunications Research Centre Vienna, Austria, and a Research Fellow at the University of Surrey, UK. Anna has earned both her habilitation and her doctoral degree in Computer Science at the University of Innsbruck, and she has a university degree in Mathematics and Computer Science degree from Novosibirsk State University, Russia.

ADOPTION OF PRECISION MEDICINE; LIMITATIONS AND CONSIDERATIONS

Nasim Sadat Mosavi and Manuel Filipe Santos

Algoritmi Research Centre, University of Minho, Guimaraes, Portugal

ABSTRACT

Research is ongoing all over the world for identifying the barriers and finding effective solutions to accelerate the projection of Precision Medicine (PM) in the healthcare industry. Yet there has not been a valid and practical model to tackle the several challenges that have slowed down the widespread of this clinical practice. This study aimed to highlight the major limitations and considerations for implementing Precision Medicine. The two theories Diffusion of Innovation and Socio-Technical are employed to discuss the success indicators of PM adoption. Throughout the theoretical assessment, two key theoretical gaps are identified and related findings are discussed.

KEYWORDS

Precision Medicine, Adoption, Artificial Intelligence, Healthcare Big Data, Open data exchange, Genomes, Biological indicators, Standards, Internet of Things, Blockchain.

1. INTRODUCTION

With the availability of healthcare big data and technological advancement, clinical practice going beyond the "one-size-fits-all" approach. Where emerging the new clinical decision-making, minimizes medical errors, cuts the cost of overtreatment, increases the quality of services offered by care providers, and saves more lives[1].

Precision medicine is an emerging approach in medical decision-making that takes into account individual genetic profile, environmental and lifestyle indicators. In 2015 the former president Obama, launched the Precision Medicine Initiative (PMI) aiming to improve the tailoring of the treatment based on individual variables[2]. PMI intends to motivate individuals to cooperate as co-researcher to manage their health by sharing their health data (e.g., genomic, genetic, longitudinal health information) throughout a trustful partnership platform, where, this cooperation results in obscuring the boundary between health and disease [3].

Projecting this approach requires early diagnosis, prevention, and tailor the treatment for the individual patient. Based on that, sharing data and linking individual patient variables to health records are expected to lead to the right drug, at the right dose to the right patient [4]. Hence, successful adoption of PM has been found in tailoring treatment to a particular patient subgroup with common molecular characteristics [5].

For the clinical application of precision medicine to be able to fit the patient with matching treatment modalities, the incorporation of various heterogeneous parameters is required. Although research is underway around the world to speed up progress on precision medicine

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 13-24, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110302

projection, and there are research programs such as the "All of Us Research"that provide innovation opportunities to address the limitations and solutions[3], still, there has not been a clear protocol and business model for emerging the PM in healthcare.

The need for collecting individual patient data (e.g., genomes, biological indicators, demographics, administration) and integrating them into the Electronic Health Records (EHRs), identifying suitable approaches and techniques to deal with big data, transforming large, multimodal data into Machine Learning (ML) algorithms for decision-making [6],[7], dealing with data protection security and ethical issues of sharing such sensitive data, adopting standards for data exchange, empowering stakeholders with education, and proposing effective pipelines for restructuring the policies and regulations are some of the considerations for employing PM.

This study aimed to propose the major limitations and considerations for adopting PM in healthcare. The paper first explains two theories, which have been used widely by literature to address remarkable indicators for successful implementation of Information Systems (IS). After that, the main limitations and challenges for projecting PM are highlighted and finally, the paper is ended by a discussion throughout examining both theories in implementation of PM. In this assessment, we identified two considerable limitations and gaps, which are not taken into the account by both theories: DoI and Socio-tech.

2. THEORETICAL FOUNDATION

The aspects of adoptinginnovation/ technology have been studied for over 30 years, and literature has used multiple theories to address the success indicator for adopting new technology. The theoretical basement of this study highlights the general consideration of Information Systems (IS) adoption through the Socio-technical theory and Diffusion of Innovation (DoI). Where both theories have been widely used in scientific research from a broad variety of disciplines [8].

2.1. Socio-Technical Theory

According to figure 1, the socio-technical theory emphasizes that the information system contains two major interrelated subsystems: the technical and social. Whereas technical aspects address tangible factors such as technology and tasks required to convert system inputs into out- puts, social aspects mostly concentrate on organizational aspects, people and the perfect harmony between technological tools and human activities are the keys for successful adoption [9],[10].

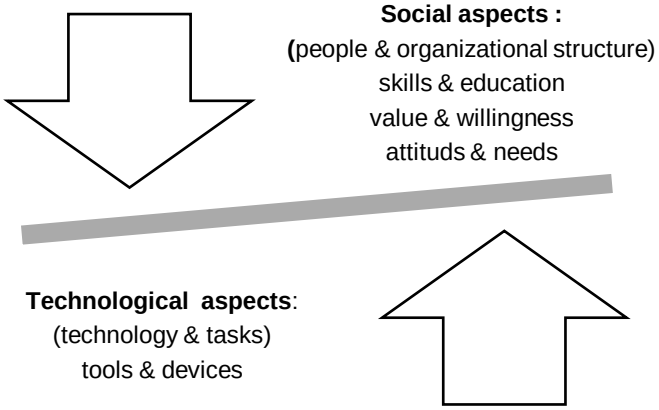


Figure 1. Socio-Tech Paradigm

2.2. Diffusion of Innovation Theory

From IS perspective, Diffusion of Innovation (DoI) expresses major indicators for successful adoption: compatibility (degree in which the new technology is compatible with the existing system)**,** technical complexity(how easy/difficult is the technology to understand)**,** the relative advantage of adopting the technology (the degree to which an innovation adds value to an existing system), triability (to what extent the innovation can experiment before adoption), and observability (how the innovation provides tangible results)[9],[11].

Roger defines diffusion as the process that innovation is communicated via specific channels over time and through the member of a social platform; he used the word "technology" and "innovation" interchangeably [8]. Thus, not only technological aspects of innovation affect a successful adoption, but also, social platform and communication including, people, policies, regulations, and management influence this process.

3. PRECISION MEDICINE

Healthcare and medicine industries are facing dramatic changes under the influence of Artificial Intelligence (AI) techniques, Business Analytics (BA). The huge amount of data is challenging the business model of empirical medicine [1], where, answering the questions such as why does a drug work for some patients and be less effective on others? Why does medicine cause side effects on some individuals? Why do cancers influence some people and do not others? need The medical practice of tailoring each patient as a unique case.

On January 20, 2015, former President Barack Obama announced the great potential health improvement to science. The "Precision Medicine" launched by him, has been a remarkable research opportunity in a new area of medical practice for improving public health; *"Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes and to give all of us access to the personalized information we need to keep ourselves and our families healthier."*; Hence the new approach to medical practice includes two components; a short-term which concentrates on cancer diseases and the long-term that considers other illnesses [4]. According to the U.S. National Library of Medicine, "Precision Medicine" is an emerging approach that considers individual differences such as genes, environment, and lifestyle for preventing and treating particular diseases [4], [12]. Furthermore, the institute of US National Cancer defines Precision Medicine as "a form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease" [13]. Moreover, the approach of PM is the category of medical sciences that aims to predict the possibility of developing a disease, achieve a precise diagnosis, and optimize the best performance of treatment for a particular patient [14]. In other words, according to figure 2, precision medicine intends to consider individual patient variables in terms of genetic, lifestyle, and environmental effects also distinguishing patients from other patients with the same presentations for improving the clinical practice, minimizing side effects, and increase the performance of treatment [15].

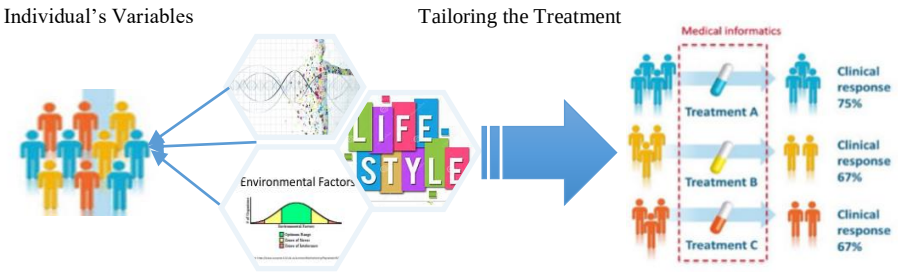


Figure 2. Precision Medicine

4. BARRIERS of PRECISION MEDICINE ADOPTION

4.1. Characteristics of the Domain

4.1.1. Adoption of Classification Systems and Clinical Terminologies

Although terminologies that contribute to PM have been built using ontologies (a form of reasoning through logics for classifications), the challenges related to the employment and integration of them in healthcare is one of the major barriers that strongly affect the development of PM application [16], [17].

Lack of nation-wide projects and less completed projects of implementing ontology-based, terminologies have caused the stakeholders to avoid solid conclusions about the projection of ontology-based terminologies. Moreover, the lack of experts in biomedical ontologies and the semantic web are another limitation for PM development. As a result, the architectures to fully integrate different types of ontologies are still immature. This lack of large developments resultsin a high risk when designing EHR systems that use multiple ontologies since a large number of terminologies and ontologies have been developed in parallel by different bodies, thus making ontology mapping among them extremely challenging[18], [19].

4.1.2. Limitations in Clinical Evidence, Outcomes, and Value Assessment Practice

Whereas, technologies are employed to develop the new approach of Medical Decision Making; PM, insufficient information, successful clinical evidence, and unavailability of universally adopted data models for validating the practical work have not guaranteed the value and outcome of PM [16].

4.1.3. Adoption of Standards for Data Collection and Integration- Clinical Data Exchange

There have been various limitations and challenges in laboratory, medication, diagnosis, radiology, pathology, clinical evidence & outcomes, and procedures, based on the availability of different data structures and related challenges of adopting the existing standards for data exchange. For example, incorporation of genetic results into EMR in a searchable way. Moreover, in many cases, tests, which are conducted in external labs, cannot be integrated into other systems. Differences between lab code systems and identifications or missing the genetic info in EHRs data are some of the key challenges related to the data exchange [16]. Furthermore, identifying the right moment for clinical data exchange becomes a competitive process that needs attention [20].

The researchers have identified the need of extending the scope of the existing standards rather than inventing a new one where data standardization for integration and exchange is required for the correct interpretation of the data elements.

Although various initiatives have worked in this line to facilitate the adoption of the data standards especially in the omics discipline such as BioSharing, (works to ensure the standards are searchable and informative.), to extend the existing standards, the relevant stakeholders should cooperate to educate the potential adopters for understanding and using the existing data standards [21].

4.1.4. Data Processing and Storage – Handling Big Data

The speed of generating a massive amount of data from various resources has not been balanced with big data management yet. The analysis and processing of such data need hardware and software facilities, which is beyond the local infrastructures of many small laboratories. Particularly, molecular and omics data analysis requires powerful computational tools [21].Whereas componential resources and related facilities require a high cost of maintenance and development [6], adopting technologies such as cloud computing and the Internet of Things (IoT)can be an alternative to deal with the high cost of storage and computational requirements. For example, Siemens Healthcare, Philips Healthcare, and GE Healthcare are the three big medical imaging companies that pioneered switching computer-intensive image processing to the Health Cloud ecosystem. Together they have developed a strategy for establishing an "All-in-one Health Cloud" to provide the unique cloud platform [2], [22]. Although cloud computing can be a potential solution to deal with big data management issues, the need for protecting healthcare data and applying security and privacy levels will be increased[6], [23]. To deal with this requirement, solutions such as access control models and blockchain technology are offered. However, each solution carries specific limitations, which need to be addressed. For example, access control models, which perform based on the access authorizations and defined regulations, are less effective against internal risk attacks. Hence, in many cases, hybrid approaches are applied to integrated access control with some other methods such as encryption. Moreover, Blockchain technology is another alternative for secure data management in cloud platforms where the possibility to delete or modify stored data is almost zero. This characteristic of blockchain criticizes the regulations of personal data privacy; when it is obligated to delete the personal data based on individual patient's rights. Hence, in this case, blockchain technology requires to solve these types of challenges through new policies, privacy models, and further studies[24].

4.1.5. Ethical and Legal Literature for Open Data Exchange

Data protection, security, and related ethical issues are another challenge. Personal health data is sensible and valuable. It has been estimated that the healthcare domain is 200 % more likely to experience a data breach than other industries. Moreover, because data generated modular via the internet and in many cases stored offline, ethical consideration prevents an open data exchange [6]. For example, unintended access (the misuse of information gained through unauthorized access) is a serious challenge when it reaches to debating about ethical issues [25].

Due to the reforming of the medical domain as the data-sharing community, the patient's role as the owner and manager of her/his health data became more valid. Thus changing the power between care providers and care receiver is considerable needs restructuring regulations and policies. On one hand, the patient is requested to share personal information, on the other hand, it gives the right to the individual to become a member of the data-sharing community thus the

roles, responsibilities, and authorities need restructuring and ethical issues should be addressed in an effective way [3].

4.1.6. Features of Healthcare Data

For PM, the major data include EHRs and Omic. Big Omic data contains a comprehensive catalog of molecular profiles (e.g., genomic, transcriptomic, epigenetic, proteomic, and metabolomics) and EHRs contain structured and unstructured clinical data. Both Omic and EHRs are challenging for analytics because of the data frequency, quality, dimensionality, and heterogeneity. Moreover, EHRs include structured data (e.g., ICD-9 diagnosis codes, administrative data, charts, and medication) and unstructured (e.g., clinical notes). Structured data includes two classes: administrate data, which remain unchanged during the entire clinical process. For example, demographics are administrative data. The second class is ancillary clinical data; the data, which frequently or continuously recorded during the clinical process such as medications and lab tests of blood pressure monitoring via sensors. This heterogeneity of data is a considerable challenge for data analysis[26]. Furthermore, in omic data, the combination of biological, instrumental, and environmental factors affects the quality of data. In EHRs, the quality of data depends on the missing data and incorrect records, which are the consequences of the clinician's teamwork to enter data, miss-interpretation, and the organization of documents for data entry. Hence, any mistakes and misinterpreting will cause wrong conclusions and inaccurate decision-making [26].

4.1.7. Adoption of Genomics in EHR (eMERGE)

Adoption of PM requires the integration of EHRs and Genomics (eMERGE). In a clinical setting, EMR (Electronic Medical Records) is used for clinical decision support and the integration between genomic/genetic data and EMR identifies causal genomic variants and genotype-phenotype associations into the EMR system. The major challenge of such integration is finding the most suitable method for storing and reprocessing the variants present in an individual or even family and the next generation in EMR. Moreover, for integrating genetic/genetic data into the EHRs, the size of genetic/genomic laboratory test results and the limited capacity of EHRs also has been identified as significant limitations. Because each individual has millions of variants and the variation of genomic data cannot fit into the current design of EHRs. One potential solution is to archive raw data in separate data repositories to be accessible once it is requiring. However, storing such unstructured data makes the processing speed slow. Another alternative is to strategize the external genetic/genomic data warehousing. In this approach, data is stored external to EHRs and the link connect and integrate particular record to related EHRs record [27].

In the addition, the genetic records stored in EHRs need to be interpreted throughout the clinical terms and vocabularies. To tackle this challenge, a couple of solutions have been suggested such as using rule-based decision support systems and using visualization elements for better presentation.

Besides, EHRs include medical records of all participants and hospitals, while EMR contains limited data for local clinics and hospitals. Therefore, standardizing the data exchange and defining protocols for better interoperability is useful for effective treatment [26]. Moreover, in addition to the technological challenges for adopting genomic/genetic in EHRs, the environmental aspects such as ethical and legal boundaries for data sharing and such integration need attention[17], [28]. To support the adoption of PM, various countries propose practical policies to protect individual's genomic/genetic data from discrimination to assure the patient about the security and privacy level of their data [29].

4.1.8. Research and Practice

Many research works have discussed the effective role of AI/ML for early diagnosis and better treatment, but also the relevant challenges have been addressed. Since projecting such approaches is still in the development stage, the lack of understanding of AI and ML is a limitation, and identifying the most effective approaches for pioneering valid clinical practice of PM needs more scientific research and practice [16]. Furthermore, the techniques for automating data collection, analysis, and processing are usually projected locally, and transferring the success story to other healthcare providers is not easy [6].

To speed up the delivery of precision medicine, more research is needed, particularly in the following biomedical big data areas and omics data integration. The big omics data analysis, for example, provides a holistic view for analyzing the patient condition and for effective prediction [26]. in many cases, applied research environments and academia have faced limitations and challenges to accessing healthcare data, in terms of privacy and security. Thus, Protected Health Information (PHI) and lack of trust are some of the critical issues that affect the development of research works in this area [16].

The second area in which research and practice demonstrate the influential effect is Patient segmentation based on similarities. Classifying patients based on biological similarities on their profile is a crucial phase in developing PM. Therefore, data mining applied in EHRs us clinical indicators such as drug responses, physiological signals, and disease susceptibility for patient classifications. However, high patient variability for disease and also the fact that many subgroups of the disease have not been identified yet, affect the performance of the practical PM application. Hence, this gap needs systematic research for validating the classifications of the patient based on EHR mining [26].

Finally, as was discussed above, healthcare requires strong clinical evidence to analyze the validity of precision medicine in practice. This requirement is completely important for expanded clinical use [5]. Therefore, to address the limitations, it is required to develop secure research-based frameworks for efficient data collection, data integration, storage and pre-processing, de-identification to serve a large community of users, support organizational policies, and provide efficient access and connectivity [16]. Such as a new research platform that minimizes the isolation between scholar platform and clinical data [21].

4.2. Lack of a Unique Definition of Precision Medicine

Although Precision Medicine is in use as a common label, by many public funding streams (e.g., Genome England, Australian Genomics, and the Center for Personalized Cancer Treatment in the Netherlands) and private funding streams (e.g., IBM Watson), but the terminology is still evolving, and multiple terms have been used interchangeably literature has used various terms such as 'pharmacogenomics' and 'P4 medicine' (preventative, predictive, participatory, and personalized), to point out different aspects of the relevant research about precision medicine. Whereas all of these conceptualizations seem to perform in the same direction [3], the lack of the unique terminology may lead to misunderstanding by the stakeholders who have the power and responsibilities and different roles in accelerating PM adoption [5], [30].

5. EXAMINING THEORIES; DISCUSSION and FINDINGS

From the socio-tech theory perspective, in the implementation of PM application two subgroups play a critical role: technological and social. Also, the effective balance between the two aspects

grantees success. Moreover, as we discussed above, the diffusion process requires, innovation to be communicated in time through social channels. Combining the major success indicators of these two theories for examining PM application implementation, we conclude that the PM application as the innovation/technology needs to be assessed in terms of the degree of its compatibility with existing systems in healthcare (e.g., workflows, business processes, hardware/software infrastructures and resources). Moreover, the benefit and advantages of using this technology should be identified and communicated widely via a social platform, where the power and role of stakeholders' networks are remarkably important. Besides, identifying the complexity, reliability, and observability of this technology needs to be evaluated via a defined research platform where the effective cooperation between academia, the healthcare industry, and technology providers results in trust and confidence to achieve the outcome of technological evaluations.

From the social and organizational point of view, as it was discussed in both theories, management support, effective communication, educating people, pipelining strategic plans for tackling the environmental pressures such as boundaries and policies make the diffusion smooth clear, and rewarding.

The adoption of precision medicine not only will affect many health systems, but also all stakeholders. (e.g., Individual patients, care providers, policymakers, technology providers, drug developers). Therefore, each level of stakeholders carries specific preferences, definitions, and the requirement for accepting the new innovative technology (PM); proposing the value assessment of the PM can be an alternative to link stockholders and facilitate and strategies education and communication required for new technological adoption.

Although the socio-tech and DoI, both address the social, organizational, and technological indicators in total, these two theories carry gaps. The limitations are identified where the characteristics of the domain and the vital needs for research and practice are discussed. As table 1, demonstrates, without solving the problems carries by healthcare (e.g., projection of terminologies and classifications, integrating genomes in EHRs, adopting standards for data exchange, challenges about data features, data security, and ethical issues, dealing with big data, selecting the most effective approaches), the emerging of PM will not be projected successfully. Moreover, presenting evidence-based validity, value assessment, proposing the unique and fashionable term and definition for PM, and identifying the most effective techniques and tools for data analysis, data extraction, and mining to fulfill the PM approach requires research and practice. This cooperation needs inputs from other disciplines such as biomedical research, statistics, economics, and ethics, and key health stakeholders; where, in practice, the interplay between several stakeholders should be taken into the account[5], [30], and finally, empowering patients, doctors and public, as the co-researcher about PM accelerate the adoption process. Making PM comprehensible, trustworthy, and equally accessible for the population as the participatory opportunities for people to participate in the production of data for the good of medical progress. Where the new forms of participation and data control are needed to be considered[31].

Table 1. Major considerations and limitations of PM adoption

Major Limitations and Considerations of PM Adoption in Healthcare	Diffusion of Innovation	Socio-technical	Characteristics of Domain	Research &Practice
Adoption of classification systems and clinical terminologies			X	X
Limitations in clinical evidence & outcomes	X			X
Adoption of standards clinical data exchange			X	
Data processing and storage – handling big data			X	
High security, ethical and legal literature	X		X	
Features of healthcare data(format, quality)			X	
Adoption of genomics in EHRs (emerge)			X	
Selecting the most effective approaches-techniques			X	X
Patient segmentation based on similarities				X
Lack of a unique definition of PM	X			X
Management aspects Education and skills Communication Environmental pressures (policies, regulations)	X	X		X

As a result, based on identified limitations of both theories (DoI, socio-tech) in IS implementation, figure 3 presents the final influential indicators for emerging PM adoption in healthcare. Where the two new aspects: "research & practice" and "characteristics of the domain" among the technology/innovation aspects and social/organizational elements are introduced.

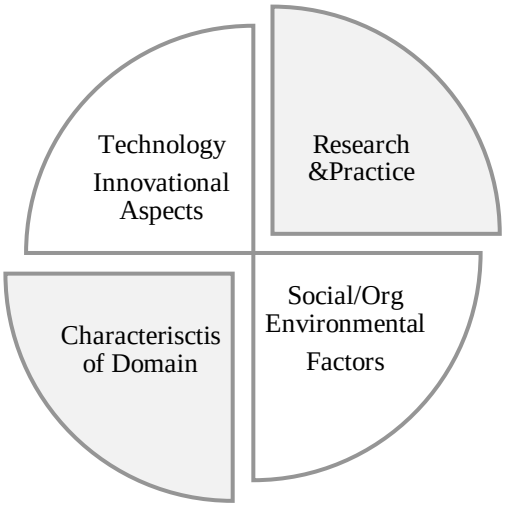


Figure 3. Success Indicators for PM Implementation- Theory Assessment

In summary, it can be concluded that PM is applicable to be introduced as socio-tech /DoI phenomena. Therefore, for integrating PM in healthcare, not only managerial and organizational

aspects needs to be addressed, but also the technical issues such as the unique definition of PM, reliability, and complexity needs considerations.

Besides, there are major limitations, which are directly related to the features of the healthcare ecosystem itself. As is mentioned above, since healthcare is a data-driven platform, therefore, limitations and challenges carry by the characteristic of data (eg., sensitive, fragmented, heterogeneous, unstructured, high volume) strongly influence this transition. For example, we discussed the possibility of employing the cloud computing solution to deal with big data management in terms of storage management and maintenance, it is also discussed the limitations that cloud computing presents in terms of security and data protection. Moreover, where blockchain technology can deal with security and data protection, the role of individuals as the owner of data and their willingness for data sharing, open another limitation. Besides, not only data characteristics are a matter but also the complexity of the domain with interconnected entities provide complicated regulations that need attention. For example, projecting PM in healthcare needs a trustful cooperation platform for data sharing by empowering individuals to participate in their health management purpose. However, such an arrangement demands a deep restructuring of standards and policies about security and data protection.

6. CONCLUSIONS

This paper discussed the key limitations and considerations for emerging Precision Medicine in healthcare. Although there have been various terms used interchangeably as PM, it is introduced as a new approach in medical decision making where the individual patient variables in genetics, lifestyle, and environmental are taken into the account. This study employed the socio-tech and DoI paradigms to explain the various technological (e.g., complexity, compatibility, trialability, tangibility, perceived value), social-organizational (communication, education, cooperation, management), and environmental (e.g., policies, regulations, ethical issues) indicators which influence the successful implementation. By examining the combined theories (socio-tech, DoI), the PM application is presented as a socio-tech and DoI phenomena through this theoreticalassessment, we identified two critical aspects, which are missed to address by both theories: “characteristics of healthcare” and “research and practice under a cooperative model of partnership and trust”. Accounting for these two identified aspects enables us to deal with the limitations and difficulties we have faced in a complex domain like healthcare. Besides, to facilitate the transition through valid clinical evidence and value assessed outcomes of PM practice. Successful adoption of PM in healthcare contributes to maximizing the quality of treatment and saving more lives by minimizing medical errors, decreasing the risk of over-treatment, minimizing the side-effect of medicines, and identifying the patient at risk for a particular disease.

ACKNOWLEDGMENTS

The work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Projects Scope: DSAIPA/DS/0084/2018.

REFERENCES

[1] M. M. Arash Shaban-Nejad, *Precision Health and Medicine : A Digital Revolution in Healthcare*. USA: Springer,2020.

[2] D. Ivanova and P. Borovska, “Internet of Medical Imaging Things and Analytics in Support of Precision Medicine for the Case Study of Thyroid Cancer Early Internet of Medical Imaging Things and Analytics in,” no. August, 2018.

[3] M. W. Vegter, "Towards precision medicine; a new biomedical cosmology," *Med. Heal. Care Philos.*, vol. 21, no. 4, pp. 443–456, 2018.
[4] H. V. Francis S. Collins, "A commentary on 'A new initiative on precision medicine,'" *Front. Psychiatry*, vol. 6, no. MAY, p. 88, 2015.
[5] E. Faulkner *et al.*, "Being Precise About Precision Medicine: What Should Value Frameworks Incorporate to Address Precision Medicine? A Report of the Personalized Precision Medicine Special Interest Group," *Value Heal.*, vol. 23, no. 5, pp. 529–539, 2020.
[6] W. Walter, N. Pfarr, M. Meggendorfer, P. Jost, T. Haferlach, and W. Weichert, "Next-generation diagnostics for precision oncology: Preanalytical considerations, technical challenges, and available technologies," *Semin. Cancer Biol.*, no. October, 2020.
[7] G. Onder, R. Bernabei, D. L. Vetrano, K. Palmer, and A. Marengoni, "Facing multimorbidity in the precision medicine era," *Mech. Ageing Dev.*, vol. 190, no. April, p. 111287, 2020.
[8] I. Sahin and F. Rogers, "Detailed Review of Rogers ' Diffusion of Innovations Theory and Educational Technology-Related Studies Based on Rogers '," vol. 5, no. 2, pp. 14–23, 2006.
[9] N. Foshay and C. Kuziemsky, "Towards an implementation framework for business intelligence in healthcare," *Int. J. Inf. Manage.*, vol. 34, no. 1, pp. 20–27, 2014.
[10] S. Šajeva, "The analysis of key elements of socio-technical knowledge management system.," *Econ. Manag.*, no. 2007, pp. 765–774, 2010.
[11] A. Turan, A. Ö. Tunç, and C. Zehir, "A Theoretical Model Proposal: Personal Innovativeness and User Involvement as Antecedents of Unified Theory of Acceptance and Use of Technology," *Procedia - Soc. Behav. Sci.*, vol. 210, no. December 2015, pp. 43–51, 2015.
[12] B. Mesko, "The role of artificial intelligence in precision medicine," *Expert Rev. Precis. Med. Drug Dev.*, vol. 2, no. 5, pp. 239–241, 2017.
[13] M. Haque, T. Islam, M. Sartelli, A. Abdullah, and S. Dhingra, "Prospects and challenges of precision medicine in lower-and middle-income countries: A brief overview," *Bangladesh J. Med. Sci.*, vol. 19, no. 1, pp. 32–47, 2020.
[14] J. Awwalu, A. G. Garba, A. Ghazvini, and R. Atuah, "Artificial Intelligence in Personalized Medicine Application of AI Algorithms in Solving Personalized Medicine Problems," *Int. J. Comput. Theory Eng.*, vol. 7, no. 6, pp. 439–443, 2015.
[15] J. L. Jameson and D. L. Longo, "Precision medicine - Personalized, problematic, and promising," *N. Engl. J. Med.*, vol. 372, no. 23, pp. 2229–2234, 2015.
[16] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Q. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, pp. 1–35, 2020.
[17] E. J. Cicali *et al.*, "Challenges and lessons learned from clinical pharmacogenetic implementation of multiple gene–drug pairs across ambulatory care settings," *Genet. Med.*, vol. 21, no. 10, pp. 2264–2274, 2019.
[18] K. Malm-nicolaisen, L. Marco-ruiz, and K. Malm-nicolaisen, *Ontology-based terminologies for healthcare*. 2017.
[19] M. A. Haendel, C. G. Chute, and P. N. Robinson, "Classification, Ontology, and Precision Medicine," *N. Engl. J. Med.*, vol. 379, no. 15, pp. 1452–1462, 2018.
[20] R. Ashkenazy, "Precision medical communication to optimize stakeholder information exchange: A '4M-Quadrant' approach," *Drug Discov. Today*, vol. 21, no. 7, pp. 1039–1041, 2016.
[21] M. Afzal, S. M. Riazul Islam, M. Hussain, and S. Lee, "Precision medicine informatics: Principles, prospects, and challenges," *arXiv*, pp. 13593–13612, 2019.
[22] M. J. Khoury, G. L. Armstrong, R. E. Bunnell, J. Cyril, and M. F. Iademarco, "The intersection of genomics and big data with public health: Opportunities for precision public health," *PLoS Med.*, vol. 17, no. 10, pp. 1–14, 2020.
[23] Y. A. Qadri, A. Nauman, Y. Bin Zikria, A. V. Vasilakos, and S. W. Kim, "The Future of Healthcare Internet of Things: A Survey of Emerging Technologies," *IEEE Commun. Surv. Tutorials*, vol. 22, no. 2, pp. 1121–1167, 2020.
[24] C. Esposito, A. De Santis, G. Tortora, H. Chang, and K. K. R. Choo, "Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy?," *IEEE Cloud Comput.*, vol. 5, no. 1, pp. 31–37, 2018.
[25] C. M. Hammack, K. M. Brelsford, and L. M. Beskow, *Thought Leader Perspectives on Participant Protections in Precision Medicine Research*, vol. 47, no. 1. 2019.

[26] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "-Omic and Electronic Health Record Big Data Analytics for Precision Medicine," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 263–273, 2017.
[27] A. N. Kho *et al.*, "Practical challenges in integrating genomic data into the electronic health record," vol. 15, no. 10, 2013.
[28] J. G. Chase *et al.*, "Next-generation, personalised, model-based critical care medicine: A state-of-the art review of in silico virtual patient models, methods, and cohorts, and how to validation them," *Biomed. Eng. Online*, vol. 17, no. 1, pp. 1–29, 2018.
[29] M. H. Ullman-Cullere and J. P. Mathew, "Emerging landscape of genomics in the electronic health record for personalized medicine," *Hum. Mutat.*, vol. 32, no. 5, pp. 512–516, 2011.
[30] R. W. Barker, "Is precision medicine the future of healthcare?," *Per. Med.*, vol. 14, no. 6, pp. 459–461, 2017.
[31] E. C. Winkler and B. M. Knoppers, "Ethical challenges of precision cancer medicine," *Semin. Cancer Biol.*, no. xxxx, 2020.

AUTHORS

Manuel Filipe Santos received his Ph.D. in Computer Science (Artificial Intelligence) from the University of Minho (UMinho), Portugal, in 2000. He is an associate professor with habilitation at the Department of Information Systems, UMinho, teaching undergraduate and graduate classes of Business Intelligence and Decision Support Systems. He is the head of Intelligent Data Systems lab and the coordinator of the Information Systems and Technology group (www.algoritmi.uminho.pt) of the R&D ALGORITMI Centre, with the current research interests: Business Intelligence; Intelligent Decision Support Systems; Data Mining and Machine Learning (Learning Classifier Systems); and Grid Data Mining. He is part of the steering committees of the master's course in Engineering and Management of Information Systems and the Doctoral Program in Information Systems and Technology.

Nasim Sadat Mosavi is a Ph.D. student at the University of Minho (UMinho), Portugal. She also works as a researcher at Centro algoritmi, (UMinho). Her research interest is Intelligent Decision Support Systems (IDSSs) using Machine learning and optimization techniques. Healthcare/Medicine is her research interest domain. Nasim graduated in Computer Science (associate's degree) and Computer Engineering-software (bachelor's degree) from the Islamic Azad University of Tehran-Iran and she pursued her master's degree in International Business from the University of Wollongong. She was involved, in the development and implementation of more than 100 successful IS projects in different positions with different industries in Dubai, Iran, K.SA.

DEEP LEARNING SELF-ORGANIZING MAP OF CONVOLUTIONAL LAYERS

Christos Ferles, Yannis Papanikolaou,
Stylianos P. Savaidis and Stelios A. Mitilineos

Department of Electrical and Electronics Engineering, University of West Attica, Aegaleo, Attica, Greece

ABSTRACT

The Self-Organizing Convolutional Map (SOCOM) combines convolutional neural networks, clustering via self-organizing maps, and learning through gradient backpropagation into a novel unified unsupervised deep architecture. The proposed clustering and training procedures reflect the model's degree of integration and synergy between its constituting modules. The SOCOM prototype is in position to carry out unsupervised classification and clustering tasks based upon the distributed higher level representations that are produced by its underlying convolutional deep architecture, without necessitating target or label information at any stage of its training and inference operations. Due to its convolutional component SOCOM has the intrinsic capability to model signals consisting of one or more channels like grayscale and colored images.

KEYWORDS

Deep Learning, Unsupervised Learning, Convolutional Neural Network (CNN), Self-Organizing Map (SOM), Clustering.

1. INTRODUCTION

Probably the most common bottleneck encountered in many deep learning approaches like Convolutional Neural Networks (CNNs) is the requirement for big labeled datasets. Constructing these datasets is a costly time-consuming procedure that frequently might end up proving infeasible for various reasons. The obvious answer to this problem is devising deep learning models that can be trained with unlabeled/uncategorized data, in other words, invent unsupervised learning algorithms for such deep networks. Aligned with this ongoing research direction one can trace a number of works that combine or hybridize Self-Organizing Maps (SOMs) with CNNs.

The gamut of these approaches –including the present one– is quite widespread, spanning the range from purely unsupervised learning algorithms up to semi (or even full) supervised ones, and from shallow networks up to architectures containing multiple hidden layers; for instance [1], [2], [3] and [4]. Meeting both requirements i.e. building a deep SOM and training it in a purely unsupervised way has proven to be a complex and difficult task. Only a small number of models exist that can be classified as unsupervised beyond any doubt [5], [6]and [7]. Equally few are the approaches that extend beyond the three hidden layer limit [8], [9] and [6].

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 25-32, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110303

The Self-Organizing Convolutional Map (SOCOM) is an attempt to overcome, at a certain extent, the aforementioned limitations. Its key characteristics and contributions are: (1) A deep architecture that is in position to expand beyond the trivial, and not particularity deep, three hidden layer limit. (2) An end-to-end purely unsupervised learning algorithm that does not necessitate the targets/labels of the training samples at any stage.

The organization and structure of the remainder of this paper is as follows. Section 2 presents in detail the SOCOM both architecturally and operationally, and subsequently, analyses the key components of the corresponding feed-forward and backpropagation procedures. Section 3 contains experimental (comparative) results and performance evaluations with different algorithms. Also, (practical) application issues are discussed in this section. In section 4, a summary is given and conclusions are drawn. Finally, section 5 gives hints for future work and suggestions for potential expansions.

2. SOCOM PROTOTYPE

A generic and at the same time characteristic SOCOM architecture consisting of multiple convolutional, pooling, fully-connected and self-organizing layers is illustrated in Figure 1. The basis of the mathematically expressed algorithmic learning procedures is presented in the following subsection. This section describes the main functionality and key methods of the SOCOM from a more macroscopic operational point of view.

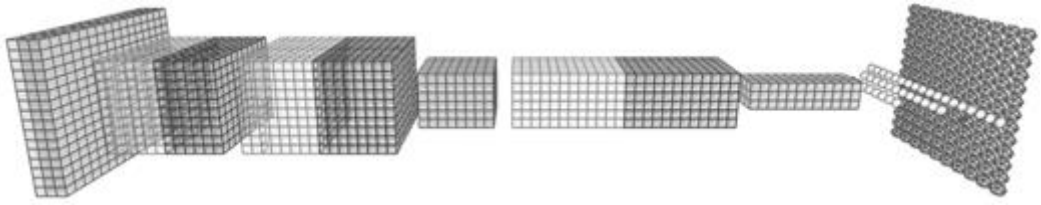


Figure 1. Detailed architecture of a SOCOM paradigm consisting of the following layers: input → convolutional→ReLU → convolutional → ReLU→pooling → convolutional → ReLU → pooling → fully-connected → fully-connected → fully-connected →output neural map.

The input layer of the SOCOM accepts any type of numerical data arranged in vectors, matrices (e.g. grayscale images) or volumes (e.g. colored images or successive images that exhibit a spatiotemporal correlation). The explicit assumption of CNNs that the inputs are images, something that makes the information propagation more efficient to implement and hugely reduces the network's parameter count, still holds in the SOCOM paradigm but does not a priori exclude all other types of input data.

As can be seen, a SOCOM comprises of a sequence of different layers with adjustable parameters. Each respective layer transforms one volume of activations to another via a differentiable function, thus facilitating the use of backpropagation during training. Stacking these layers in series eventually forms a full SOCOM architecture (Figure 1).

Similarly to other CNNs the convolutional layer consists of a set (or bank) of tunable filters/kernels. Despite of its usually small size, every filter extends through the full depth of the input volume. During the forward propagation each filter slides along the width and height of the input volume by performing convolutions. Strictly mathematically speaking the convolution operation carried out here is the same as cross-correlation except that the kernel is rotated by 180^0. In the long run this procedure yields a two-dimensional activation map that contains the responses of the respective filter at each spatial position. The hypothesis (which is currently

backed up by several experimental findings in the literature) is that the network will tune its filters so that they activate when they trace some type of visual feature, edge, or pattern. Stacking the activation maps generated by the respective layer's bank of filters produces the activation volume (or feature map) that is fed to the following (hidden) layer. As has been discussed, the units in a layer are only connected to a small region of the layer before it. This underlying weight sharing strategy, which is the aftereffect of using small filters, ends up reducing the overall number of trainable weights hence introducing sparsity, and at the same time, making the architecture suitable for manipulating images.

Neural networks' essential characteristic of nonlinearity (frequently in the form of the sigmoidal or hyperbolic tangent functions) is retained in CNNs by applying element-wise a non-saturating function to the activation volume of the preceding convolutional layer. The norm, that also the SOCOM adheres to, is to apply the rectified linear unit (ReLU) function to each individual activation produced by the convolutional layer. It has been shown, that such nonlinearities result in richer and more elaborate representations along the network architecture.

At certain points in the convolutional-ReLU layer hierarchy a pooling layer is inserted. Essentially, pooling performs a downsampling operation solely along the width and height spatial dimensions of the input volume. The reduction of such blocks of activations to just a single value has several positive aftereffects: (1) the number of parameters and related computations is reduced, (2) sparseness is introduced, and (3) overfitting is avoided.

After several convolutional and pooling layers, it is common to transition to fully-connected layers where the high-level abstract representations are formed. These densely connected layers are identical to the layers of the standard multilayer neural network. The first fully-connected layer decomposes the activations of its input volume into a one-dimensional vector and connects them to every unit it has. Subsequent layers consist of units which receive all the activations from the previous layer and perform a dot product followed by a nonlinearity. Fully-connected layers are not spatially arranged anymore something that prohibits the use of convolutional layers after a fully-connected layer.

Finally, a SOM lattice of topologically arranged neurons acts as the output layer. Each of its neurons receives the activations of every unit in the last fully-connected layer. The magnitude of each neuron's activation is based on a distance metric between the input activations and its codebook parameters. The neural mapping of the input image coincides with the position of the neuron that produces the optimal fit with respect to the computed activations and the neighborhood kernel (which has been defined over the topology of the neural grid). Apart from mapping this particular type of nonlinear projection can be further exploited for data clustering and visualization.

It is also interesting to note that the proposed SOCOM architecture is in position to incorporate any number of layers (from the previous types) in any permutation. There are only two limitations: (1) after the first fully-connected layer convolutional layers cannot be used, (2) the output layer needs to be a SOM grid.

2.1. Forward Propagation

As has been demonstrated a generic SOCOM architecture consists of an input layer, L hidden layers (convolutional, ReLU, pooling and fully-connected ones) and an output layer (viz. lattice of ordered neurons). The novel component of the SOCOM is its neural output map and in particular the different from the norm energy function that is associated with it.

The output layer that consists of G topologically arranged neurons performs a mapping of its input representations onto its neural map. More specifically, the projection of an input representation on the SOCOM plane is defined as the neuron yielding the lowest weighted squared Euclidean distance between the last hidden layer's outputs $o_i^L$and its corresponding codebook parameters $u_{g,i}$ where weighting refers to the neighborhood kernel/function $h_{e,g}$ defined over the topology of the neural grid. Frequently, this neuron (denoted as c) is referred to as "winner". Algorithmically, this best-matching winner neuron is given by:

$$c = arg\min_{e} \sum_{g=0}^{G-1} h_{e,g} \sum_{i=0}^{P-1} \left(o_i^L - u_{g,i}\right)^2 \quad \text{(eq. 1)}$$

where$P$ is the total number of units in the last hidden layer L. Additionally, this particular type of nonlinear projection can be further exploited for data clustering and visualization procedures.

2.2. Backpropagation

The purpose of being in positon to compute an error or loss function is dual. First, a definite quantification/assessment of the network's performance is obtained. Second, learning takes place via the optimization of the network's weights to minimize this specific error. This error function can be a number of different things, such as binary cross-entropy or sum of squared residuals. Differently from supervised approaches, learning in the case of SOCOM does not necessitate any type of desired or target values at any stage; thus giving rise to a pure unsupervised deep learning algorithm. The corresponding error/cost/loss function (or alternatively, the penalty term) is symbolized as E and is defined as:

$$E = \sum_{c=0}^{G-1} N(c) \sum_{d=0}^{G-1} h_{c,d} \frac{1}{2} \sum_{i=0}^{P-1} \left(o_i^L - u_{d,i}\right)^2 \quad \text{(eq. 2)}$$

Where

$$N(c) = \begin{cases} 1, & c = arg\min_{e} \sum_{d=0}^{G-1} h_{e,d} \sum_{i=0}^{P-1} \left(o_i^L - u_{d,i}\right)^2 \\ 0, & otherwise. \end{cases} \quad \text{(eq. 3)}$$

For gradient descent backpropagation the updates that need to be performed are for the weights, the biases and the deltas. The utilized energy formula by the SOCOM is in accordance with the variation proposed in [10]and has been applied in a number of hybrid SOM networks [11], [12].

3. EXPERIMENTS

The experimental investigation strategy that has been followed serves a dual purpose. First, a (mainly quantitative) comparison against a comprehensive series of similar/related deep SOMs is achieved. These models cover the full range of SOMs that extend beyond the mainstream two layer architecture (a single input layer connected to an output neural map) by employing at least one intermediate hidden layer between their inputs and outputs. Second, the conducted experiments act as a proof of concept for the proposed network by tangibly demonstrating/verifying its capabilities and clustering performance, given the modelling problem under consideration.

Following the justified requirement of comparing the SOCOM approach with an as wide as possible gamut of likewise SOM approaches the MNIST benchmark choice was unavoidable since: (1) the landslide of published deep SOMs report results (frequently, exclusively only) on the MNIST dataset, (2) it is traditionally the entry point dataset of experimental investigation when it comes to testing deep learning algorithms. The MNIST benchmark used in the current experimental setup is Yann LeCun's version [13] which contains handwritten numerical digits that have been size-normalized and centered in a fixed-size image. The dataset consists of 60000 training examples and 10000 testing examples; it is an almost balanced collection where the highest deviating category (in terms of sample size) is the handwritten digit "1" (approximately 11.2% in the train set and 11.4% in the test set instead of the expected 10%).

Before moving further an important point should be made. An end-to-end purely unsupervised learning algorithm that does not necessitate the targets/labels of the training samples at any stage. If these are provided they can potentially be used, but, typically, an unsupervised model should be is in position to function even when these are absent or missing. Nevertheless, there is merely a handful of approaches that adhere to strict unsupervised training criteria [5], [6], [7]and the ones reporting results on the MNIST database are specifically indicated in Table 1. Frequently, in the literature, an "unsupervised" model with a supervised or self/semi-supervised training procedure is proposed. Apart from the fact that this defeats the purpose and it is deluding, it is practically of questionable use. If the targets/labels of the input data are utilized during training then why resort in clustering results (which are intrinsically of coarser/qualitative nature) when the alternative of classification results (which are more detailed/informative) is on the table.

Evaluating the quality of a clustering output and, in particular in the case of SOMs, of a mapping output is a non-trivial task that has been tackled by introducing various internal and external criteria. Internal criteria are more qualitative in the sense that they evaluate clustering results indirectly (e.g. by means of organization, compactness/sparseness, isolation and preservation), whereas external are more quantitative since by measuring the match between clustering and external (e.g. human-based) categorizations they are in position to provide more precise assessments. In the related literature, the most widely used external criterion, in particular for clustering tasks, is purity:

$$PUR = \frac{1}{S}\sum_{p=1}^{P}\max_{1\le t\le T}\left|s_p \cap s_t\right|. \qquad \text{(eq. 4)}$$

The subscript p denotes the partitioning of a set of S samples into P distinct clusters (a posteriori estimated by the model); similarly, the subscript t denotes the assignment of these samples into T categories (a priori defined in the dataset). As expected its resulting values lie in the $[0, 1]$ interval. Obviously purity identifies with accuracy given that the majority voting principle is utilized for labeling each individual cluster. Although purity intuitively is rather straightforward/precise it tends to favor small (in sample numbers) clusters like singletons.

On a related note, a distinction should be drawn between obtaining accuracies with a posterior labeling of neurons (based on data labels) and obtaining accuracies with the addition of a supervised model/layer (like MLP, SVM or fully-connected Softmax network). Obviously, the latter approaches' results are misleading since the unsupervised networks' outputs are treated as input features to a supervised network (which is obviously trained in a supervised manner). This type of experimental testing does reveal characteristics of the unsupervised module's output feature space but is by no means indicative of the network's clustering capabilities and performance.

The SOCOM architecture that has been utilized in the present series of experiments closely follows that of resnet18 [14] upon which appropriate modifications have been carried out. Specifically, the first hidden 2D convolutional layer has been replaced by a 2D convolutional layer that accepts single channel signals/images, the last fully-connected layer has been removed, an output layer implementing the neural map has been added, followed by an 1D pooling layer for facilitating the backpropagation optimization algorithm. Standard stochastic gradient descent backpropagation with momentum [15] is used for training the network. Transfer learning [16]is also utilized for obtaining the initial weight/parameter values of the hidden layers that are shared with the resnet18 architecture. The codebook parameters have been initialized according to the methodology described in [17], using a uniform distribution. The lower and upper limits of the value ranges used for the learning rate and momentum hyper-parameters have been estimated according to the technique described in [18]. Output neurons are arranged onto a 2D hexagonal grid; the Gaussian neighborhood kernels' standard deviations start with a value equal to half the largest dimension on the grid and decrease linearly to one map unit, during training. The performance (in terms of accuracy) and main characteristics of a list of indicative deep SOMs including SOCOM are summarized in Table 1.

Table 1. The architectural/algorithmic characteristics of various deep SOMs and their respective accuracies on the MNIST dataset.

Model/Network	Accuracy (%)	End-to-End Unsupervised Learning	Number of Layers
(Aly, 2020)[4]	99.43		3
(Braga, 2020)[19]	98.36		4
SOCOM	97.35	●	20
(Wang, 2017)[9]	96.7		8
(Liu, 2015)[2]	96.17		3
(Friedlander, 2018)[5]	87.7	●	3
(Wickramasinghe, 2019)[3]	87.12		2
(Wickramasinghe, 2018)[20]	84.87		2

As can be seen the proposed SOCOM outperforms the majority of previous approaches by utilizing a purely unsupervised learning algorithm which is capable of handling (through the backpropagation of gradients) all the necessary computations needed for adjusting the underlying deep architecture. All the rest of the approaches, apart from [5], use extensively label/target information throughout their training procedures for reaching the reported accuracy rates. This observation further demonstrates the capabilities of the SOCOM since it is in position to perform better against (or almost at par with) models that access richer information like the label/class information of input images of handwritten digits. It should be noted that by taking into consideration the other purely unsupervised deep SOM the SOCOM achieves nearly 10% improved accuracy. Last, it is also important to reiterate that algorithmically the SOCOM model is not restricted to single channel input signals (like the grayscale MNIST images) but it is capable of incorporating three channel inputs (i.e. colored images) or input volumes of higher dimensions. This can be accomplished in a straightforward way by not replacing the first hidden convolutional layer's filter shape with the downscaled one used in these experiments.

4. SUMMARY

One of the central dogmas in the field of machine learning (which differently from dogmas in other domains, is continuously being backup up experimentally) is that the stratification of

several levels of nonlinearity is the key to tackle complex recognition tasks, infer higher-level correlations between variables and representations of data, and, in general, mimic and model the way human perception and ingenuity function. SOCOM aligns with the ongoing research towards combining nonlinearities of neurons into networks for modelling highly complex and increasingly varying functions. It is doing this by trying to remain loyal to the unsupervised learning guidelines of necessitating as less label information as possible.

It has been shown, both algorithmically (i.e. in theory) and experimentally (i.e. in practice), that this first working SOCOM prototype is in position to incorporate a deep architecture (evidently deeper in comparison to the deep SOMs reported in the literature) which is trained with a gradient backpropagation algorithm tailored to meet the requirements of the architecture's complexity, depth and parameter size. As has been discussed previously, the proposed algorithm not only is along the lines of the optimization methods which are proven to work with deep networks but also keeps the required label/target information to a minimum. Further, due to the fact that the first hidden layers of SOCOM's architecture are convolutional, the data that can be modelled are not restricted to grayscale images (i.e. single channel ones) but instead can consist of an arbitrary number of channels e.g. colored images (i.e. three channels) or even sequences of images/signals; such data rarely can be processed by the currently published deep SOMs.

5. FUTURE WORK

It is reasonable that this proof-of-concept study of the SOCOM prototype could give rise to a number of closely-related research directions pointing towards expanding and enriching the model, and towards making full use of its clustering capabilities in real-world complex problems. More specifically, an omnidirectional research plan could involve: (1) The construction of deeper SOCOMs based for instance on the resnet34, resnet50 and resnet152 architectures [14]. (2) Gradually utilizing the backpropagation flow of gradients in adjusting/tuning layers further deep down the architecture. (3) Incorporating diverse deep network configurations that are based upon other well-known paradigms like Alexnet[21], VGG [22], and GoogLeNet[23]. (4) In depth and in detail analysis and evaluation of the various optimization methods provided by the Pytorch framework. (5) Using existing deep learning visualization techniques up to the last hidden representation layer, and, subsequently, treating visualizations as "inputs" to the ordered neuron output array. The final objective in this case is having either a visualization of what the map models/clusters [24] or a projection of the achieved higher-level representations onto the output map.

ACKNOWLEDGEMENTS

This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020" in the context of the project "Self-Organizing Convolutional Maps" (MIS 5050185).

The authors would like to thank the anonymous reviewers for their constructive comments and insightful remarks.

REFERENCES

[1] Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks, 8(1), 98-113.

[2] Liu, N., Wang, J., & Gong, Y. (2015, July). Deep self-organizing map for visual classification. In 2015 international joint conference on neural networks (IJCNN) (pp. 1-6). IEEE.

[3] Wickramasinghe, C. S., Amarasinghe, K., & Manic, M. (2019). Deep self-organizing maps for unsupervised image classification. IEEE Transactions on Industrial Informatics, 15(11), 5837-5845.
[4] Aly, S., &Almotairi, S. (2020). Deep Convolutional Self-Organizing Map Network for Robust Handwritten Digit Recognition. IEEE Access, 8, 107035-107045.
[5] Friedlander, D. (2018). Pattern Analysis with Layered Self-Organizing Maps. arXiv preprint arXiv:1803.08996.
[6] Pesteie, M., Abolmaesumi, P., &Rohling, R. (2018). Deep neural maps. arXiv preprint arXiv:1810.07291. maps.
[7] Stuhr, B., &Brauer, J. (2019, December). CSNNs: Unsupervised, Backpropagation-free Convolutional Neural Networks for Representation Learning. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 1613-1620). IEEE.
[8] Part, J. L., & Lemon, O. (2016, October). Incremental on-line learning of object classes using a combination of self-organizing incremental neural networks and deep convolutional neural networks. In Workshop on Bio-inspired Social Robot Learning in Home Scenarios (IROS), Daejeon, Korea.
[9] Wang, M., Zhou, W., Tian, Q., Pu, J., & Li, H. (2017, October). Deep supervised quantization by self-organizing map. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1707-1715).
[10] Heskes, T. (1999). Energy functions for self-organizing maps. In Kohonen maps (pp. 303-315). Elsevier Science BV.
[11] Ferles, C., Papanikolaou, Y., & Naidoo, K. J. (2018). Denoisingautoencoder self-organizing map (DASOM). Neural Networks, 105, 112-131.
[12] Ferles, C., &Stafylopatis, A. (2013). Self-organizing hidden markov model map (SOHMMM). Neural networks, 48, 133-147.
[13] LeCun, Y., Bottou, L., Bengio, Y., &Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[15] Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013, May). On the importance of initialization and momentum in deep learning. In International conference on machine learning (pp. 1139-1147). PMLR.
[16] Sharif Razavian, A., Azizpour, H., Sullivan, J., &Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 806-813).
[17] Glorot, X., &Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings.
[18] Smith, L. N. (2017, March). Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV) (pp. 464-472). IEEE.
[19] Braga, P. H., Medeiros, H. R., &Bassani, H. F. (2020, July). Deep Categorization with Semi-Supervised Self-Organizing Maps. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
[20] Wickramasinghe, C. S., Amarasinghe, K., Marino, D., & Manic, M. (2018, July). Deep self-organizing maps for visual data mining. In 2018 11th International Conference on Human System Interaction (HSI) (pp. 304-310).
[21] Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997.
[22] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
[24] Ferles, C., Beaufort, W. S., &Ferle, V. (2017). Self-Organizing Hidden Markov Model Map (SOHMMM): biological sequence clustering and cluster visualization. In Hidden Markov Models (pp. 83-101). Humana Press, New York, NY.

TOWARDS COMPARING MACHINE LEARNING MODELS TO FORESEE THE STAGES FOR HEART DISEASE

Khalid Amen[1], Mohamed Zohdy[1], and Mohammed Mahmoud[2]

[1]Department of Electrical and Computer Engineering,
Oakland University, Rochester, MI, USA
[2]Department of Computer Science and Engineering,
Oakland University, Rochester, MI, USA

ABSTRACT

With the increase in heart disease rates at advanced ages, we need to put a high quality algorithm in place to be able to predict the presence of heart disease at an early stage and thus, prevent it. Previous Machine Learning approaches were used to predict whether patients have heart disease. The purpose of this work is to compare two more algorithms (NB, KNN) to our previous work [1] to predict the five stages of heart disease starting from no disease, stage 1, stage 2, stage 3 and advanced condition, or severe heart disease. We found that the LR algorithm performs better compared to the other two algorithms. The experiment results show that LR performs the best with an accuracy of 82%, followed by NB with an accuracy of 79% when all three classifiers are compared and evaluated for performance based on accuracy, precision, recall and F measure.

KEYWORDS

Machine Learning (ML), Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbors (KNN).

1. INTRODUCTION

1.1. Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) that is increasingly utilized within the field of heart disease medicine. It is essentially how computers make sense of data and decide, or classify, a task with or without human supervision. The conceptual framework of ML is based on models that receive input data (e.g., images or text) and through a combination of mathematical optimization and statistical analysis predict outcomes (e.g., favorable, unfavorable, or neutral) [2]. We have used five ML algorithms in our previous work to predict multiple stage heart disease. The first one is SVM, it can recognize non-linear patterns for use in facial recognition, handwriting interpretation or detection of fraudulent credit card transactions. So-called boosting algorithms used for prediction and classification have been applied to the identification and processing of spam email.

The second algorithm is Random Forest (RF), it can facilitate decisions by averaging several nodes. The third algorithm is Gradient Tree Boosting (GTB), which is a ML technique for regression and classification problem that produces a prediction model in the form of an ensemble of weak prediction models. The fourth algorithm is Extra Random Forest (ERF), it is an

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 33-44, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110304

ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. The fifth algorithm is Logistic Regression (LR), it is a classification algorithm, that is used where the response variable is categorical [3]. The idea of LR is to find a relationship between features and probability of particular outcome. We have previously described technical details of each of these algorithms, and found out that LR is the best algorithm in terms of accuracy, precision, recall and F measure to predict multiple stages of heart disease. We have also used several tools and methods such as Python libraries, graphs, and Pseudocodes to test the performance of each algorithm. In this paper, we are going to implement two more algorithms, Naïve Bayes (NB) and K-Nearest Neighbors (KNN) to predict multiple stage heart disease and compare with our winning algorithm, LR [1].

Machine Learning in healthcare is becoming more widely used and is helping patients and clinicians in many different ways. ML can play an essential role in predicting presence/absence of locomotor disorders, heart diseases and more. ML, when applied to health care, is capable of early detection of disease, which can aid to provide early medical intervention. Heart disease predication has been a very hot topic for ML, for example, the analysis of heart disease has become vital in health care sectors. The success of ML in the medical industry is its capability in analyzing the huge amount of data gathered by the health sector and its effectiveness in decision-making [4] [5].

As we have used in our previous work to conduct this prediction, a Jupyter notebook was constructed in Python using the publicly available Cleveland dataset for heart disease, which has over 300 unique instances with 76 total attributes. From these 76 attributes, only 14 of them are commonly used for research to this date. In addition, the libraries and coding packages used in this analysis are: SciPy, Python, NumPy, IPython, Matplotlib, Pandas, ScikitLearn and Scikit-Image.

1.2. Heart Disease

Heart disease is the major cause of morbidity and mortality globally and accounts for more deaths annually than any other cause. According to the World Health Organization (WHO), an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three-quarters of these deaths took place in low and middle-income countries [6] [14].

Heart disease is the number one killer of both men and women. Heart disease can happen at any age, but the risk increases as people get older. Children of parents with heart disease are more likely to develop heart disease themselves. African-Americans have more severe instances of high blood pressure than Caucasians and a higher risk of heart disease. Heart disease risk is also higher among Mexican-Americans, American Indians, native Hawaiians and some Asian-Americans. This is partly due to higher rates of obesity and diabetes. Genetic factors likely play some role in high blood pressure, heart disease and other related conditions [7] [8].

The silver lining is that heart attacks are highly preventable and simple lifestyle modifications (such as reducing alcohol and tobacco use, eating healthily and exercising) coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high-risk patients because of the multi-factorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where ML and data mining come to the rescue [12].

2. RELATED WORK

In our previous work [1] and research, we have implemented five algorithms SVM, RF, GTB, LR and ERF to predict multiple stage heart disease using the Cleveland dataset. We concluded that the LR algorithm performed better in terms of accuracy, precision, recall and F measure as shown in table 1:

Table 1. Algorithms comparison

Algorithm	Accuracy	Precision	Recall	F Measure
SVM	80%	91%	78%	84%
LR	82%	91%	80%	85%
RF	77%	89%	76%	82%
GTB	74%	80%	76%	78%
ERF	79%	89%	78%	83%

Additionally, many researchers have completed a lot of work on data analysis and survivability analysis through ML and Data Mining (DM) approaches [7].

In [8], [10] the author applied Decision Tree (DT), LL, NB, SVM, KNN, PCA, ICA classifier respectively to analyze kidney disease data. Early detection and treatment of the diseases prevents it from getting to the worst stage, making it not only difficult to cure, but also impossible to provide treatment. Breast cancer affects many women, so researchers work on different classifiers such that DT, SMO, BF Tree and IBK help to analyze the breast cancer data and examine the performance of the related techniques in order to accurately predict breast cancer using DT and Weka software. RBF Network, Rep Tree and Simple Logistic DM techniques are used to predict and resolve the survivability of breast cancer patient. Simple Logistic is used for dimension reduction and proposed RBF Network and Rep Tree model used for fast diagnosis of the other diseases [19] [20] [21].

3. BACKGROUND OF CLEVELAND DATASET

In our previous work, we have used the Cleveland dataset for multiple stage heart disease prediction. We plan to use the same dataset to compare the results of the new algorithms we are adding with LR. Experiments with the dataset have concentered on 14 attributes that were used. The list is in Table 2.

Table 2. Cleveland dataset attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1 = true 0 = false

Restecg	Discrete	Resting Electrocardiograph
Thalach	Continuous	Exercise Max Heart Rate Achieved
Exang	Discrete	Exercise Induced Angina: 1=yes 0=no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1=up sloping 2=flat 3=down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that range between 0 and 3
Tha	Discrete	3=normal 6=fixed defect 7=reversible defect
Class	Discrete	Diagnosis classes: 0=No Presence 1=Least likely to have heart disease 2=>1 3=>2 4=More likely have heart disease

The database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by DL researchers to date. The "num" field in the figure refers to the presence of heart disease in the patient. It is integer valued from zero (no presence) to four. Experiments with the Cleveland database have concentrated on attempting to distinguish presence (values 1,2,3,4) from absence (value 0) [16].

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

4. BACKGROUND ON LOGISTICS REGRESSION ALGORITHM

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In LR the dependent variable is always binary. LR is mainly used for prediction and calculating the probability of success [17] [18]. An LR model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. The weaknesses are that LR cannot properly deal with the problems of non-linear and interactive effects of explanatory variables. LR is a regression method for

predicting a dichotomous dependent variable. In producing the LR equation, the maximum likelihood ratio was used to determine the statistical significance of the variables. LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a LR model but is suited to models where the dependent variable is dichotomous [22]. The LR model for p independent variables can be written as: [1]

$$H(Y=1) = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp)}} \quad (1)$$

where $P(Y=1)$ is the probability of the presence of CAD and $\beta 0,\ \beta 1,\ \beta 2, \ldots, \beta p$ are regression coefficients. There is a linear model hidden within the logistic regression model. The natural logarithm of the ratio of $P(Y=1)$ to $1-P(Y=1)$ gives a linear model in Xi: [2]

$$g(x) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) \quad (2)$$

$$= \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp$$

The $g(x)$, has many of the desirable properties of the LR model. The independent variables can be a combination of continuous and categorical variables.

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there will only be two possible classes [22] [23]. In simpler words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a LR model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, diabetes prediction, cancer detection etc. [22] [23].

4.1. Type of Logistics Regressions

Logistic Regression means binary LR having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, LR can be divided into following types [15] [24]:

- Binary or Binomial – In such a classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.
- Multinomial – In such a classification, dependent variable can have three or more possible unordered types, or the types having no quantitative significance. As an example, these variables may represent “Type A” or “Type B” or “Type C”.
- Ordinal – In such a classification, dependent variables can have three or more possible ordered types, or the types having a quantitative significance. For example, these variables may represent “poor”, “good”, “very good” or “excellent” and each category can have scores such as 0, 1, 2 or 3.

5. METHODOLOGY

The proposed methodology using two classification techniques; NB and KNN. We use these two classifications to predict heart disease as the proposed methodology shown in Fig 2. These classifiers are used to improve the prediction. We applied the classifiers in Fig 5 to heart disease data that comes from the Cleveland dataset to predict in which of five stages a patient has heart problems. The performance of these classifiers are to evaluate on the bases of accuracy, precision recall and F measure, then we compare the results of these classifications with LR in terms of accuracy, precision recall and F measure.

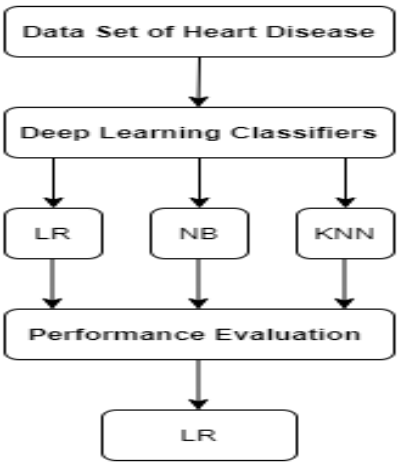


Figure 1: Proposed Methodology

The dataset of heart is taken from Machine LearningRepository UC Irvine, the classifier taking it as input for disease prediction. These classifiers are implemented in Python language. Python is a powerful interpreter language and a reliable platform for research [25]. The accuracy of prediction increased by comparing the results of these five classifiers using evaluation parameters. The experimental result describes which classifier is best between them.

5.1. Evaluation Parameters

- Accuracy is defined as the number of accurately classified instances divided by the total number of instances in the dataset as in (3).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

- Precision is the average probability of relevant retrieval as described in (4).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

- The recall is defined as the average probability of complete retrieval as defined in (5).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

- F- Measure is calculated by using both precision and recall as shown in (6).

$$\text{F Measure } = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (6)$$

Some evaluation parameters in DM are accuracy, precision, recall and F measure. Where True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [14].

Where all evaluation parameters accuracy, precision, recall and F measure are calculated from dataset when splitting the dataset into training data and test data. The Pseudocodes for the evaluation parameters are as follows:

```
Def evaluationParameters(X_train, y_train, X_test, y_test):
X_train ← fit_transform(X_train)
Classifier ← sklearn()
y_pred ← classifier.predict(X_test)
cm_test ← confusion_matrix(y_pred, y_test)
y_pred_train ← classifier.predict(X_train)
    cm_train ← confusion_matrix(y_pred_train,    y_train)
    training_accuracy=(cm_train[0][0]+cm_train[1][1])/  len(y_train)
    test_accuracy=(cm_test[0][0]+cm_test[1][1])/len(y_test)
training_percision = cm_train[0][0]/(cm_train[0][0] + cm_train[1][0])
test_percision=cm_test[0][0]/(cm_test[0][0]+cm_test[1][0])
training_recall = cm_train[0][0]/(cm_train[0][0] + cm_train[0][1])
test_recall = cm_test[0][0]/(cm_test[0][0] + cm_test[0][1])
training_f_measure ← (2 * training_percision * training_recall)/(training_percision +
training_recall))
test_f_measure ← (2 * test_percision * test_recall)/(test_percision + test_recall))

return (training_accuracy, test_accuracy, training_percision, test_percision, training_recall,
test_recall,  training_f_measure,  training_f_measure)
```

6. DATASET

To perform the research, the heart disease dataset is used. This heart disease dataset contains 14 attributes and 303 instances. This dataset is taken from UCL repository. It's an online repository that contains 412 diverse datasets. UCI provides data for ML to perform analysis in a different direction. The UCI database is known for its extensiveness in data, its completeness and accuracy [23].

6.1. Machine Learning Classifiers:

In this continuous research, two additional classification methods are implemented in python using the pandas and keras libraries. These models are used to improve prediction. These classifiers are compared with LR to find out which of the five stages best predicts the chance of heart disease in patients. In the next section, we briefly describe these classification techniques/classifiers.

1) Naïve Bayes (NB): are probabilistic classifiers based on Bayes theorem with naïve independence assumption between the predictors or features. NB classifier assumes, that the existence of a particular feature is not related to the existence of any other feature in a class [26].

For example, apples are considered a fruit, if it is red and round. Even if these are features related to each other or depend upon the presence of other features. NB classifier consider all these features to contribute independently to probability identifying that the fruit is an apple.

NB model is easy to construct and particularly valuable for large data sets and a Bayes theorem gives a way to calculate posterior probability P(c|x) from likelihood (predictor probability) P(x|c), class prior probability P(c) and predictor prior probability P(x) as shown in (7)

$$p(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (7)$$

2)K-Nearest Neighbors (KNN): is used for regression and classification problems. KNN is commonly used for classification problems [26] [27]. KNN classifier store all the existing cases then classified new cases by the majority votes of its neighbors. The case is assigned to that class, which is most common to its k nearest neighbors, measured by distance function. These distance functions are Euclidean Eu, Manhattan Ma and Minkowski Mi calculated using (8), (9) and (10) respectively.

$$Eu = \sqrt{\sum_{k=1}^{n} (pk - qk)^2} \quad (8)$$

$$Ma = \sqrt{\sum_{k=1}^{n} |pk - qk|} \quad (9)$$

$$Mi = \left(\sqrt{\sum_{k=1}^{n} |pk - qk|} \right)^{1/r} \quad (10)$$

Whereas r is the parameter, n is the number of attributes or dimensions. pk and qk are respectively, the kth element of objects p and q [28].

6.1.1. Scaling Data

To accomplish the five stages output prediction for a patient to be diagnosed with one of five stages, it is important to scale the data so the machine learning algorithms do not overfit to the wrong features. Using the MinMaxScaler() method on Python, the values are scaled per features based on the minimum and maximum between 0 and 1. This keeps the information from being lost but allows the machine learning algorithms to correctly train with the data. The training data and test data are scaled between 0 and 1 and the output data is scaled between 0 and 1 as well. Then, the scaled output value is mapped as follows in table 3:

Table 3: Five Stages

Output Value	Stage
0	No disease presented
0 < and <= 0.25	Stage1
0.25 < and <= 0.5	Stage2
0.5 < and <= 0.75	Stage3
0.75 < and <= 1	Advance disease presented

7. EXPERIMENTAL RESULT

The experiment is conducted for the prediction of heart disease stages by applying two machine learning classifiers. From the experiment results, we have identified that LR performs better compared to the other four ML classifiers in the prediction of these diseases. In this experiment, we use multiple stages of heart disease prediction to forecast the stage at which a person is determined to have heart disease. In previous works [19] [20] [21], the study used two outcome predications, either a person has the disease or not; that is represented by (0 ,1) or (true, false). The Pseudocodes for the experiment are as follows:

```
data_frame ← read_CSV_file
X ← data_frame [column: 0 - 12]
y ← data_frame [column: 13]
target ← preprocessing.scale(y)
data ← preprocessing.scale(X)
        for k ← 0 to data - 1
        if data[k] = 0 then
        data[k] ← 'no disease'
        if data[k] > 0 && data[k] <= 0.25 then
        data[k] ← 'stage1'
        if data[k] > 0.25 && data[k] <= 0.5 then
        data[k] ← 'stage2'
        if data[k] > 0.5 && data[k] <= 0.75 then
        data[k] ← 'stage3'
        else
        data[k] ← 'disease presented'

X_train,X_test,y_train,y_test←train_test_split(X, y, test_size=0.2, random_state=0)
svm(X_train, y_train, X_test, y_test)
lr(X_train, y_train, X_test, y_test)
rf(X_train, y_train, X_test, y_test)
gtb(X_train, y_train, X_test, y_test)
erf(X_train, y_train, X_test, y_test)
```

The Figures 4, 5, 6, and 7 show the performance of various evaluation parameters in the prediction of heart disease. The experimental results show the comparison of LR, NB, and KNN classifiers and evaluate the performance on the bases of accuracy, precision, recall and F measure. In all classifiers, LR still performs the best with an accuracy of 82%, followed by NB with an accuracy of 79% and KNN with 70%. So, we can conclude that LR has better performance that ERF, GTB, SVM, RF, NB, andKNN, where LR is better than that ERF, GTB, SVM, and RF from previous evaluation, and it is better that NB and KNN in this evaluation.

8. CONCLUSIONS

The importance of extracting the valuable information from raw data has very good consequences in many fields of life such as the medical area, business area, and more. In this study, we proposed a multiple stage detection model of heart disease based on three algorithms, NB and KNN in this paper and LR from our previous work or evaluation to compare which one performs better. The proposed detection model was tested on well-known Cleveland dataset in order to provide a fair benchmark against existing studies. Based on the experimental results, our proposed model was able to outperform heart disease detection methods with respect to accuracy,

precision, recall and F measure. The result reflected the highest result obtained showed that Logic Regression has a better result comparing to the other two methods or algorithms. The performance was further enhanced using feature selection techniques. The feature selection techniques helped to improve the accuracy, precision, recall, and F measure of the ensemble algorithms. The experiment results show that LR performs the best with an accuracy of 82%, followed by NB with an accuracy of 79%, and KNN with an accuracy of 70% when all three classifiers are compared and evaluated for performance based on accuracy, precision, recall, and F measure.

ACKNOWLEDGEMENTS

This paper and the research behind it would not have been possible without the grace, the bounty, and the blessing of almighty Allah (God) first and foremost and the exceptional support of my professors, Mohamed Zohdy and Mohammed Mahmoud. Their enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track from my first encounter with machine learning research to the final draft of this paper.

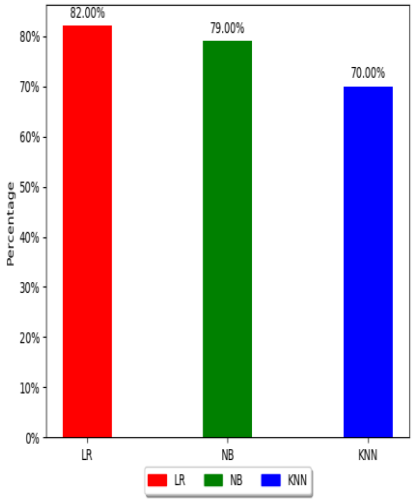


Figure 2: Heart Disease Accuracy

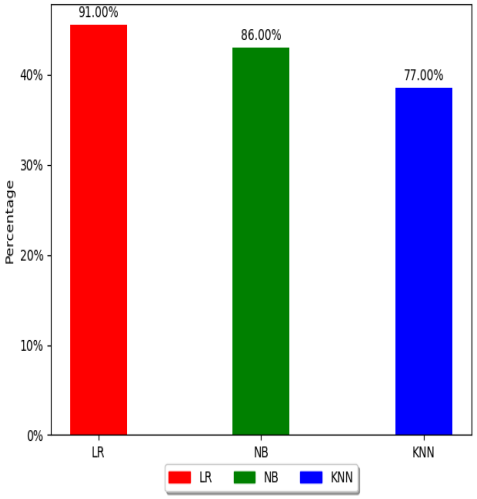


Figure 3: Heart Disease Precision

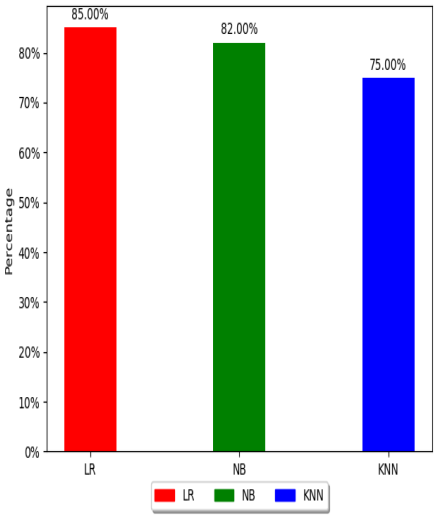


Figure 4: Heart Disease Recall

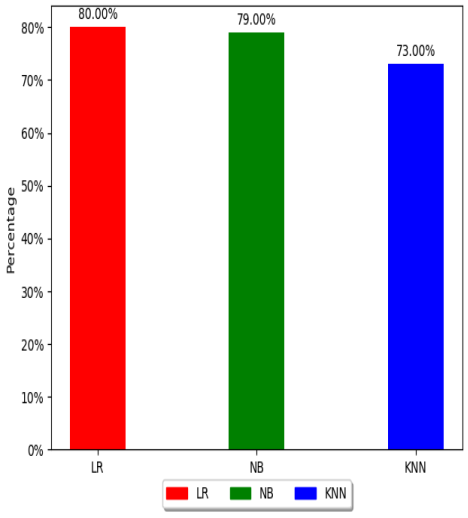


Figure 5: Heart Disease F Measure

Table 4: Five Stages

Algorithm	Accuracy	Precision	Recall	FM
LR	82%	91%	80%	85%
NB	79%	86%	79%	82%
KNN	70%	77%	73%	75%

REFERENCES

[1] K. Amen, M. Zohdy, M. Mahmoud, "Machine Learning For Multiple Stage Heart Disease Prediction". 7th International Conference on Computer Science, Engineering and Information Technology, pp. 205-223, September 26th, 2020.

[2] S. Riyaz, K. Sankhe, S. Ioannidis, K. Chowdhury, "Deep Learning Convolutional Neural Networks for Radio Identification". IEEE Communications Magazine. 56, 146–152 (2018).

[3] A Thompson, "Deep Learning on RF Data", March 29th, 2018

[4] N. Pasham, "Authenticating 'low-end wireless sensors' with deep learning + SDR", August 3rd, 2019.

[5] Z. L. Tang, S. M. Li, L. J. Yu, "Implementation of deep learning-based automatic modulation classifier on FPGA SDR platform". Electronics (Switzerland). 7 (2018), doi:10.3390/electronics7070122.

[6] Deep Learning in Healthcare, https://missinglink.ai/guides/deep-learning- healthcare/deep-learning-healthcare/

[7] Applied Deep Learning - Part 1: Artificial Neural Networks, https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

[8] Building A Deep Learning Model using Keras, https://towardsdatascience.com/building-a-deep-learning-model-using-keras-1548ca149d37

[9] B. Riyanto et al, "Software Architecture of Software-Defined Radio (SDR)", ITB Research Center on ICT, Institute Teknologi Banding, Indonesia.

[10] W. Liu et al., "Using deep learning and radio virtualization for efficient spectrum sharing among coexisting networks", Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST (Springer Verlag, 2019), vol. 261, pp. 165–174.

[11] Robert Sanders, "Distant galaxy sends out 15 high-energy radio bursts", https://news.berkeley.edu/2017/08/30/distant-galaxy-sends-out-15-high-energy-radio-bursts/, August 30th 2017.

[12] What Is Long-Term Care? https://www.nia.nih.gov/health/what-long-term-care

[13] Who Needs Care? https://longtermcare.acl.gov/the-basics/who-needs-care.html

[14] Bella Vista Health Center Blog, https://www.bellavistahealth.com/blog/2017/6/26/difference-between-short-term-care-and-long-term-care

[15] Emergency care, https://www.oregonlaws.org/glossary/definition/emergency_care

[16] Heart Disease in Cleveland, https://www.rpubs.com/aepoetry/log_reg_heart

[17] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications, 0975-8887, April, 2013.

[18] A. Khemphila, V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients", IEEE CISIM, October 2010.

[19] S. A. Kaur Guneet, "Predict Chronic Kidney Disease using Data Mining Algorithms in Hadoop," international J. Adv. Comput. Eng. Netw. , vol. 5, no. 6, pp. 1–5, 2017.

[20] J. Joshi, R. Doshi, and J. Patel, "Diagnosis and Prognosis Breast Cancer Using Classification Rules," Int. J. Eng. Res. Gen. Sci., vol. 2, no. 6, pp. 315–323, 2014, [Online]. Available: www.ijergs.org.

[21] V. Chaurasia and S. Pal, "Data mining techniques: To predict and resolve breast cancer survivability," Int. J. Comput. Sci. Mob. Comput. IJCSMC, vol. 3, p. 15, 2017

[22] J. Hoffman, "Logistic regression is used for binary data", Chapter 33 - Logistic Regression, Academic Press, 2019.

[23] A. Yalcin, S. Reis, A.C. Aydinoglu, T. Yomralioglu, "A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods", Trabzon, NE Turkey, January 2011

[24] D. Speelman, "Logistic regression: A confirmatory technique for comparisons in corpus linguistics", Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy, 2014.
[25] Json Brownlee, How to Develop an Extra Trees Ensemble with Python, https://machinelearningmastery.com/extra-trees-ensemble-with-python/, Apr. 2020
[26] M. Islam, J. Wu, M. Ahmadi, M. Sid-Ahmed, Maher, "Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers", 2008.
[27] A. L. Duca, C. Bacciu and A. Marchetti, "A K-nearest neighbor classifier for ship route prediction," OCEANS 2017 - Aberdeen, Aberdeen, 2017, pp. 1-6, doi: 10.1109/OCEANSE.2017.8084635.Evaluation ofk-Nearest Neighbor classifier performance for direct marketing
[28] Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics 4 (1985): 580-585.

AUTHORS

Khalid Amen is a System Engineering and Computer Science PhD student in the Electrical and Computer Engineering department, Oakland University, Rochester, MI, USA.

Dr. Mohammed Zohdy is a professor in the Electrical and Computer Engineering department, Oakland University, Rochester, MI, USA.

Dr. Mohammed Mahmoud is a professor in the Computer Science and Engineering department, Oakland University, Rochester, MI, USA.

ROLLING BEARING FAULT DIAGNOSIS AND PREDICTION BASED ON VMD-CWT AND MOBILENET

Jing Zhu, Aidong Deng, Shuo Xue, Xue Ding and Shun Zhang

School of energy and environment, southeast university, Nanjing , China

ABSTRACT

When deep learning is used for rolling bearing fault diagnosis, there are problems of high model complexity, time-consuming, and large memory. In order to solve this problem. This paper presents an intelligent diagnosis method of rolling bearings based on VMD-CWT feature extraction and MobileNet, VMD is used to extract the signal features, and then wavelet transform is used to extract the time-frequency features. After the image is enhanced, the MobileNet network is trained. In order to accelerate the convergence speed, this paper adds transfer learning in the network training process, and migrates the weights of the first several layers pretrained to the corresponding network. Experimental results based on bearing fault data sets show that after adopting VMD-CWT, the accuracy of mobilenet increased from 68.7% to 94%, and its network parameters were reduced by an order of magnitude compared with CNN.

KEYWORDS

Mobilenet, Variational modal decomposition, Continuous wavelet transform, Rolling bearing.

1. INTRODUCTION

Rolling bearing is an important mechanical device, which has important practical significance for social and economic development[1]. However, due to the harsh operating environment and the high incidence of rolling bearing failures, rolling bearing failures often cause huge casualties and economic losses. Using machine learning and deep learning methods to carry out abnormal detection and research on rolling bearings to achieve intelligent fault diagnosis is of important practical significance for timely detection of faults, early warning and predictive maintenance, safe operation of units, improvement of unit operation efficiency, and avoidance of accidents [2-4].

V. Purashotham[5] presents a new method for detecting localized bearing defects based on wavelet transform. Bearing race faults have been detected by using discrete wavelet transform (DWT). Vibration signals from ball bearings having single and multiple point defects on inner race, outer race, ball fault and combination of these faults have been considered for analysis.Liu[6]proposes a new feature fusion method to extract new features using kernel joint approximate diagonalization of eigen-matrices (KJADE). Hoang[7] proposed a bearing fault diagnosis method based on deep convolutional neural network structure.Using vibration signal as input data directly, it has high accuracy and robustness in noise environment. Although machine learning algorithms could identify faults based on extracted features automatically, the shallow structures astrict its ability to learn more abstract and discriminative information from the input automatically.

Convolutional Neural Network has been widely used in the field of computer vision and industry for its excellent image classification effect. However, with the improvement of model classification accuracy, the depth and complexity of the model are increasing[8]. Taking the Deep Residual Network (RESNET) proposed at the end of 2017 as an example, The model size

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 45-57, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110305

of Resnet50 is 98MB, the number of model layers is 152 layers, the number of parameters is 25636712, and the top-5 Accuracy is 0.921.The model size of Inceptionresnetv2 is 215MB, the number of layers is 572, the number of parameters is 55873736, and the top-5 Accuracy is 0.953.A small increase in model precision brings about a huge increase in model parameters[9].

However, in real application scenarios, especially in the transmission system of a wind turbine, if a mechanical failure is not handled in time, it will bring serious consequences.

Fault diagnosis requires strong real-time, low latency, and fast response . The complexity of excessive model would lead to prolonged calculation time. In case of equipment failure, saving time can avoid personnel injuries and economic losses caused by larger accidents. Second, the lightweight model would reduce memory, can save server memory, reduce server size, and even the equipment of processing could evolve towards mobile or embedded devices[10].

Therefore, it is of great research significance to apply the small but efficient CNN model to the fault diagnosis of rolling bearings of wind turbines. At present, the main research direction is to train the model, and then compress the trained complex model. Or it could conduct the research in the process of model design and design smaller models for training. The aim of this method is to reduce the size of the model while improving the accuracy of the model[11]. The research idea of MobileNet is the latter, which is to build a smaller model.The size and parameters of the model are greatly reduced while the model precision is slightly reduced, the enhanced images are classified based on the MobileNet network to achieve the purpose of fault diagnosis.

At the same time, due to the high ambient noise of rolling bearings in engineering environment, this paper used the Variational modal decomposition(VMD) to denoise the signal firstly, then uses the continuous wavelet transform(CWT) to draw the time-frequency image of the signal, finally , the MobileNet network is used to classify the enhanced image.

2. MATERIALS

2.1.Continuous wavelet transform

In the field of fault diagnosis, many applications are based on feature extraction, When the unit has a fault or abnormal state, the signal is often accompanied by corresponding feature components. Therefore, the detection of these components in the signal has become an important content in fault diagnosis. For example, in the fault diagnosis of rolling bearings and gears, the appearance of periodic pulse components often indicates the occurrence of faults.Wavelet transform is the convolution operation between the original signal and the wavelet function, which actually measures the similarity degree between the signal and the wavelet function [12]. In this way, by selecting different wavelet basis functions, the content of the components that are close to the wavelet shape in the signal could be detected, which can be used to detect the characteristic components in the signal.

Expand any function f(t) in L2(R) space under wavelet basis, and call it the continuous wavelet transform(CWT) of function f(t). The wavelet transform adjusts the window size according to the frequency automatically, and could perform multi-resolution analysis. Through the wavelet time-frequency analysis, the frequency transformation law of the signal with time is obtained, which reflects the difference between the rolling bearing faults. In this experiment, CWT is used to generate a two-dimensional matrix of wavelet coefficients, and then the mobilenet is used to automatically extract the time-frequency characteristics of vibration signals. CWT have very loose requirements for wavelet basis functions, and the following conditions are sufficient:

$$C_{\varphi} = \int_{-\infty}^{+\infty} \left| \frac{\varphi(\omega)^2}{\omega} \right| d\omega < \infty \tag{1}$$

Where, ω——Angular frequency

φ(ω)——Fourier transform of wavelet function φ(t)

The definition of CWT is as follows:

$$W(a,b)=\int_{-\infty}^{\infty}x(t)\frac{1}{\sqrt{a}}\psi(\frac{t-b}{a})dt \tag{2}$$

Where, a, b——The two parameters of wavelet transform are the expansion factor and the size factor. For example, when b=5 and a=2 are given, the wavelet basis function is shifted by 5 units and reduced by 2 times.

2.2. Variational modal decomposition

The VMD decomposes the input signal x into a specified number (K) of quasi-orthogonal band-limited intrinsic mode functions (BLIMFs) μ_k with unknown but separable spectral bands. The fundamental principle of VMD can be expressed as the solving of a constrained variational problem[13]:

$$\min_{\{u_{k_1}\},\{\omega_{k_1}\}}\left\{\sum_{k_1}^{K}\left\|\partial_t[(\delta(t)+\frac{j}{\pi t})*\mu_{k_1}(t)]e^{-j\omega k_1 t}\right\|_2^2\right\} \tag{3}$$

$$s.t\quad \sum_{k_1}^{K}u_{k_1}=f$$

Where f is the original signal, μ_{k1} is the k_1 th intrinsic mode function(IMF) component and $\delta(t)$ denotes pulse signal; j is the imaginary unit, ω_{k_1} denotes the center frequency of the k_1 th IMF component; $\|\bullet\|_2$ is 2-norm.

The penalty parameter α and LaGrangian multiplier $\lambda(t)$ are introduced to solve the above-constrained issue:

$$L(\{u_{k_1}\},\{\omega_{k_1}\},\lambda(t)=\alpha\sum_{k_1=1}^{K}\left\|\partial t[\delta(t)+\frac{j}{\pi t})*u_{k_1}(t)]e^{-j\omega_{k_1}t}\right\|_2^2+\left\|f(t)-\sum_{k_1}u_{k_1}(t)\right\|_2^2$$
$$+<\lambda(t),f(t)-\sum_{k_1=1}^{K}u_{k_1}(t) \tag{4}$$

IMFs u_{k_1} and their corresponding center frequencies ω_{k_1}, the LaGrangian multiplier λ ,are subsequently updated as:

$$\hat{u}_{k_1}^{n+1}(\omega)=\frac{\hat{f}(\omega)-\sum_{i\neq k_1}\hat{u}_i(\omega)+\frac{\hat{\lambda}(\omega)}{2}}{1+2\alpha(\omega-\omega_{k_1})^2} \tag{5}$$

$$\omega_{k_1}^{n+1}=\frac{\int_0^{\infty}\omega\left|\hat{u}_{k_1}(\omega)\right|^2 d\omega}{\int_0^{\infty}\left|u_{k_1}(\omega)\right|^2 d\omega} \tag{6}$$

$$\lambda^{n+1}(\omega) = \lambda^{n} + \vartheta[f(\omega) - \sum_{k_1=1}^{K} u_{k_1}^{n+1}] \tag{7}$$

The above update process is implemented repeatedly until the following convergence criterion is satisfied:

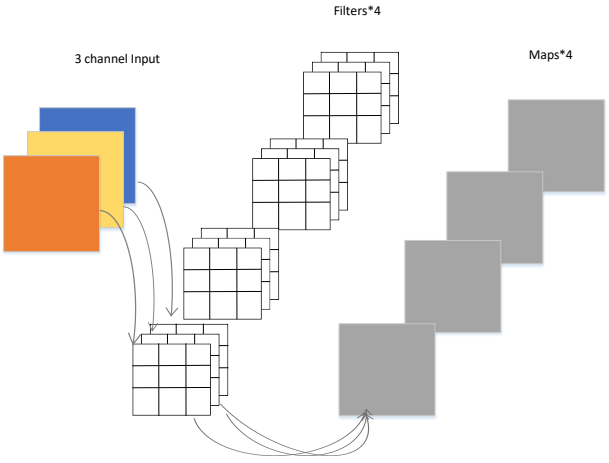
$$\sum_{k_1=1}^{K} \frac{\| u_{k_1}^{n+1} - u_{k_1}^{n} \|_2^2}{\| u_{k_1}^{n} \|_2^2} < \varepsilon \tag{8}$$

Where ε is set as 10^{-6}.

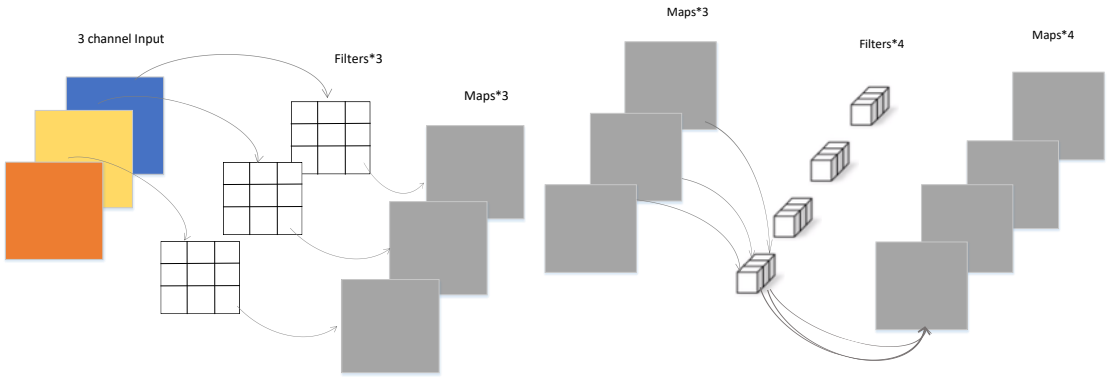
2.3. Mobilenet

Conventional Convolutional Neural Network has a large memory requirement, which makes it impossible to run on mobile devices and embedded devices, MobileNet is a lightweight CNN network proposed by Google. Compared with the traditional convolutional neural network, it greatly reduces the model parameters and calculation amount under the premise of slightly lower accuracy. The core of the model is Depthwise Convolution (DSC), which decompositions standard Convolution into depthwise(DW) Convolution and pointwise(PW) Convolution[14].

It can be seen from Figure 1 that in traditional convolution, channel of convolution kernel is equal to input eigenmatrix channel, and output eigenmatrix channel is equal to the number of convolution kernel.And in DW convolution, channel of the convolution kernel is equal to one, the input eigenmatrix channel is equal to the number of convolution cores is equal to the output eigenmatrix channel.



(a) normal convolution



(b) DW convolution (c) PW convolution

Fig 1 The structure diagram of normal convolution and depthwise Separable Convolution

Where, D_F represents the height and width of the input feature matrix, Mrepresents the depth of the input feature matrix; D_K represents the depth of the input feature matrix, Nrepresents the depth of the output feature matrix

Then, the calculation quantity of ordinary convolution is: $D_K \bullet D_K \bullet M \bullet N \bullet D_F \bullet D_F$, the calculation quantity of DW+PW is : $D_K \bullet D_K \bullet M \bullet D_F \bullet D_F + M \bullet N \bullet D_F \bullet D_F$ The ratio of calculation amount of MobileNet to that of traditional convolution is:

$$\frac{D_K \bullet D_K \bullet M \bullet N \bullet D_F \bullet D_F}{D_K \bullet D_K \bullet M \bullet D_F \bullet D_F + M \bullet N \bullet D_F \bullet D_F} = \frac{1}{N} + \frac{1}{D_K^2} \quad (9)$$

The value of N is generally large, so the result is mainly determined. At present, the size of CNN convolution kernel is generally 3*3, so theoretically, the calculation amount of ordinary convolution is about 8~9 times that of DW+PW convolution.

Where, the value of N is generally large, So the result is mainly determined by D_K , at present, the size of the CNN convolution kernel is generally 3*3. In theory, the calculation amount of ordinary convolution is about 8-9 times that of DW+PW convolution.

MobileNetV2 uses the inverted residual structure, as shown in Fig2. The inverted residual structure increases the dimensionality of the input matrix by a 1*1 convolution, and then uses a 3*3 DW convolution kernel for convolution, and then uses a 1*1 convolution kernel for dimensionality reduction[15-17].

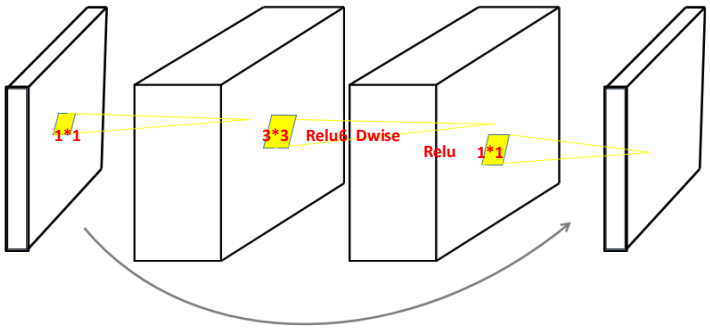


Fig 2 Inverted residual structure

The first two layers of the inverted residual structure use Relu6 as the activation function. In the last convolutional layer, a linear activation function is used instead of the ReLU activation function to avoid causing a large loss of low-dimensional feature information. Fig 3 is the main flow chart of MobilenetV2.

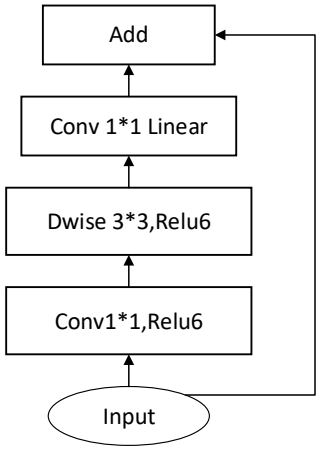


Fig 3 MobilenetV2 flowchart

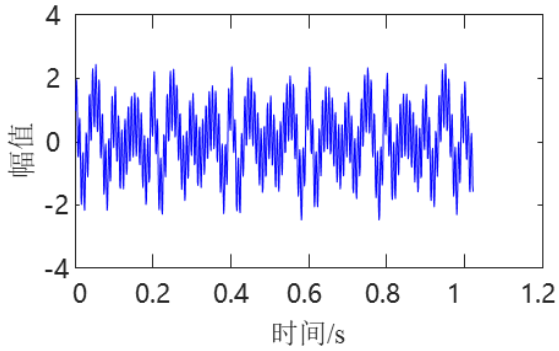
3.MAIN IDEA

3.1 Validity analysis of time-frequency graph of CWT

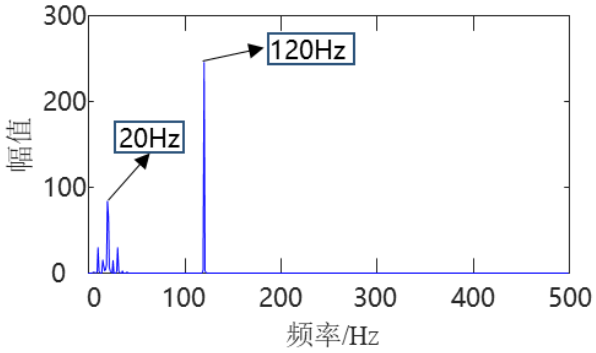
In order to prove the effectiveness of the CWT, The simulation of rolling bearing vibration signals were used to verify the results. A simulation signal with AM and FM characteristics is constructed. Time-frequency analysis based on PWVD and CWT are carried out for the simulation signal respectively. The simulation signal expression is as follows:

$$x(t) = (1+0.5\sin(2\pi \cdot 5.5t))\cos(2\pi \cdot 20t+0.8 \cdot \sin(2\pi \cdot 10t))+\sin(2\pi \cdot 120t) \quad (10)$$

Set the sampling frequency to Fs = 1000 Hz, the number of sample points N = 1024, and the waveform and power spectrum are shown in Fig4. It is known that the simulation signal is composed of an FM signal with a fundamental frequency of 20 Hz and a modulation frequency of 10 Hz, a 5.5 Hz amplitude modulation signal and a sine signal with a frequency of 120 Hz. Therefore, the 120 Hz frequency component in the signal always exists. In addition, There is also a frequency component that fluctuates with time around the fundamental frequency of 20Hz, with a frequency range of 12Hz~28Hz.These two frequency components can also be observed from the power spectrum.



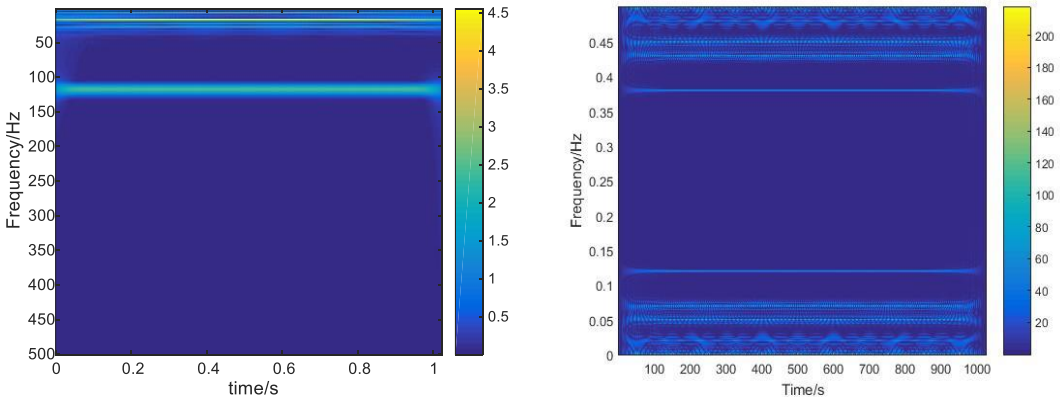
（a）Simulation signal waveform diagram



（b）Power spectrum of simulated signal

Fig 4 Simulation signal waveform and power spectrum

Time-frequency analysis based on PWVD and continuous wavelet was carried out respectively for the simulation signals. As shown in Fig5, the time-frequency diagram is obtained, in which Fig5 (a) is the time-frequency diagram of wavelet continuous transformation. It can clearly observe the frequency component that always exists at 120Hz and the frequency component that is slightly blurred around the fluctuation of 20Hz.Fig5 (b) is the PWVD time-frequency diagram, and intermittent 120Hz and 20Hz frequency components could be observed, and they are not clear. In contrast, CWT time-frequency diagram has higher time-frequency resolution and can clearly present each frequency component.



(a) CWT (b) PWVD

Fig 5 The time-frequency diagram of CWT and PWVD

3.2. Data enhancement

When classifying images with fault information, the lack of samples would lead to over-fitting of the model. Therefore, data enhancement can be performed on the images in the training set to improve the number of samples in the training set and the generalization ability of the neural network model . Image enhancement refers to generating new training samples by flipping, cropping, changing grayscale, contrast, and color. Thereby improving the invariance of the model such as scaling, and preventing the model from overfitting effectively.

3.3.Transfer learning

In order to accelerate the convergence speed of the model in the training process, the model used the transfer learning method to transfer the weights of the first several layers pretrained in other data sets of MobileNet to the corresponding network structures, and trains different neural network models on this basis.

3.4.Rolling Bearing Fault Diagnosis Based on VMD-CWT and MobileNet

In this paper, the VMD algorithm is used to extract the signal features, the CWT was used to extract the time-frequency features. After the data was enhanced, the MobileNet network was trained. In order to accelerate the convergence speed, in this paper, transfer learning was added in the network training process, and the weights of the first several layers which are pretrained on other data sets are transferred to the corresponding network.

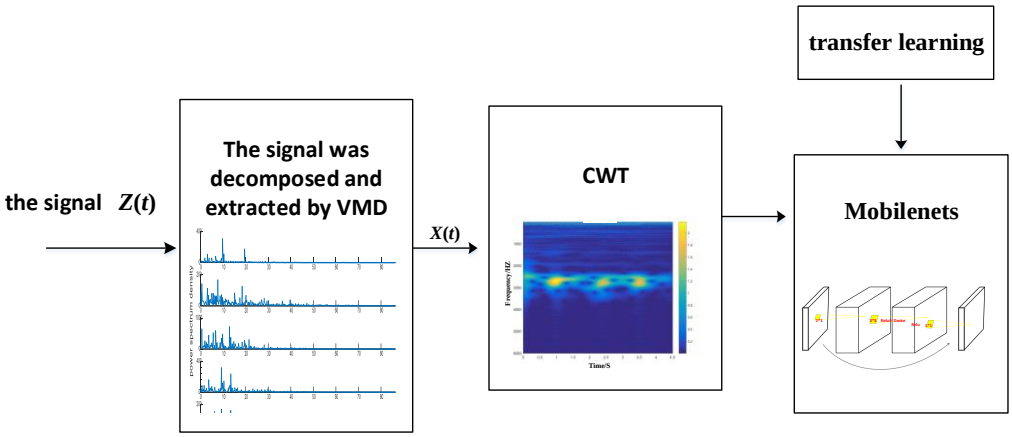


Fig 6 Flow chart of the proposed method

4.EXPERIMENTAL ANALYSIS OF PLANETARY GEAR BOX

4.1. The introduction of experimental platform

In this section, the data set of rolling bearing from Case Western Reserve University is used. Fig7 is the physical diagram and schematic diagram of the experimental platform.

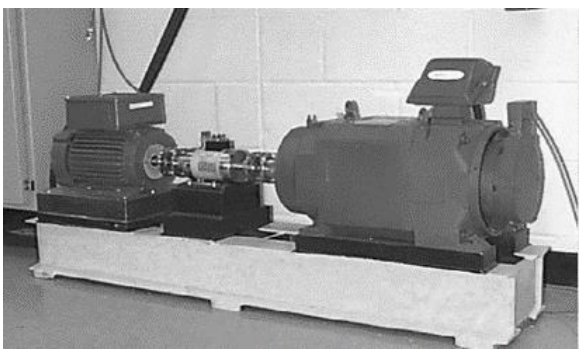


Fig 7 experimental platform

As shown in Fig7, the experimental platform consists of a 2 HP motor, torque sensor/encoder, power meter, and control electronics. They are respectively arranged on the left, middle and right side of the test bench.The test bearing supports the motor shaft, and a single point fault is introduced into the test bearing through EDM. The fault diameter is 0.007, 0.013 and 0.021 feet, all of which are SKF bearings. The specific parameters are shown in Table 1. In the experiment, an acceleration sensor was arranged to collect vibration data, and the sensor was arranged at the 12 o'clock position of the drive end and the fan end of the motor housing. The vibration signal is collected by a 16-channel DAT recorder, and later processed in the MATLAB environment. The sampling frequency of the signal sampling process is 12kHz and 48kHz.The fault of the outer ring is fixed, so the position of the fault relative to the bearing load area has a direct effect on the vibration response of the motor/bearing system.In order to make a quantitative study of this effect, the outer ring of the bearing at the drive end and the fan end were prepared with 3 o 'clock, 6 o 'clock and 12 o 'clock faults respectively in the experiment.

Table 1 The parameters of test bench bearing

Installation location	Bearing designation	outer diameter	inner diameter	• Ball diameter	contact angle	Ball number
drive end	SKF6205	52	25	7.94	0	9
Fan end	SKF6203	40	17	6.75	0	8

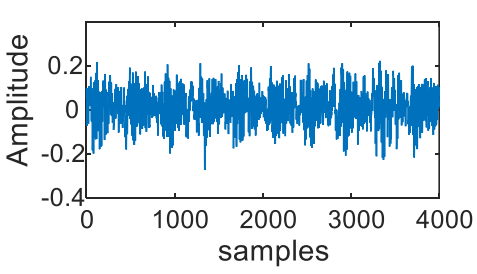
In this section, vibration signals at 1797r/min are selected to construct the data set. A total of 5 state samples are constructed, which are normal, 1 rolling body faults, 1 inner ring faults and 2 outer ring faults. The number of sampling points is 2000, and the specific number of sampling points and the categories corresponding to faults are shown in Table 2.

Table 2 Classification of rolling bearing datasets

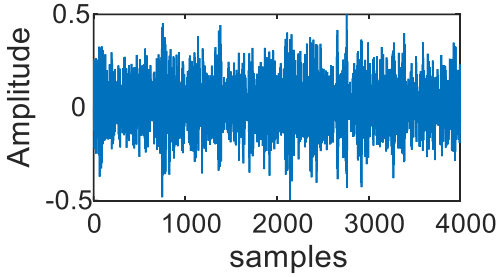
sample number	Crack diameter	fault type	label
60	0	normal	0
30	0.007	rolling body faults	1
30	0.007	inner ring faults	2
30	0.007	outer ring faults@3	3
30	0.007	outer ring faults@6	4

4.2.signal analysis

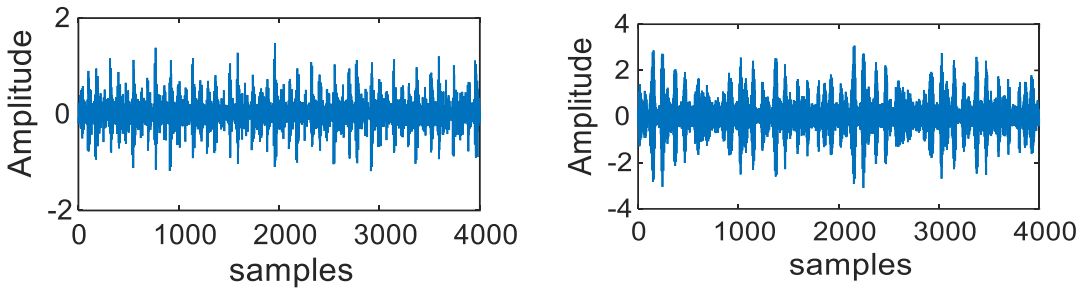
Fig8 shows the waveforms of four vibration signals of rolling bearings under the condition of 0.007 inch crack



(a) normal



(b) rolling body faults

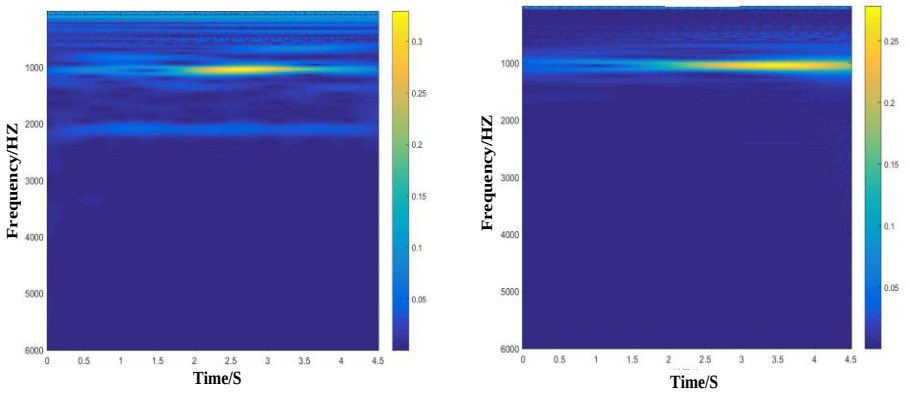


（c）inner ring faults （d）outer ring faults

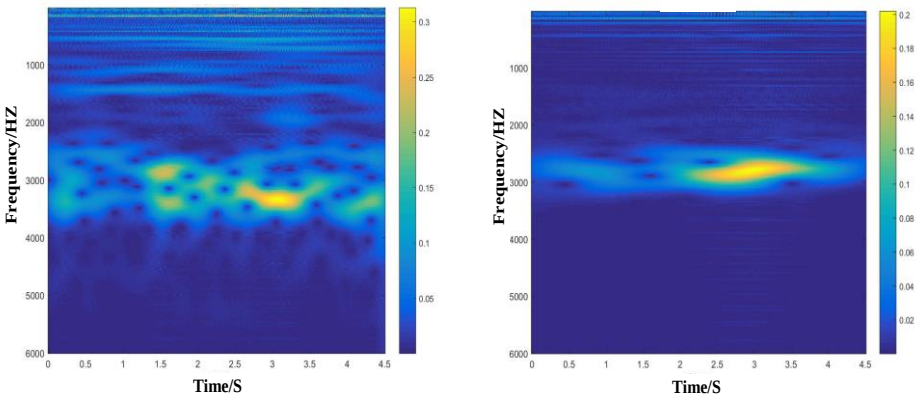
Fig 8 Time domain diagram of vibration signal

4.3.Experimental demonstration

As can be seen from the Fig8, the amplitude of different states varies greatly, the amplitude of normal state is the smallest, and the amplitude of outer ring fault is the largest. Among them, the impact of outer ring and inner ring faults is obvious. In order to improve the recognition rate of time-frequency images, VMD feature extraction was first carried out on the signals. The following is the comparison of time-frequency images before and after VMD feature extraction under normal and rolling body fault conditions.



(a) Normal-before (b) Normal-After



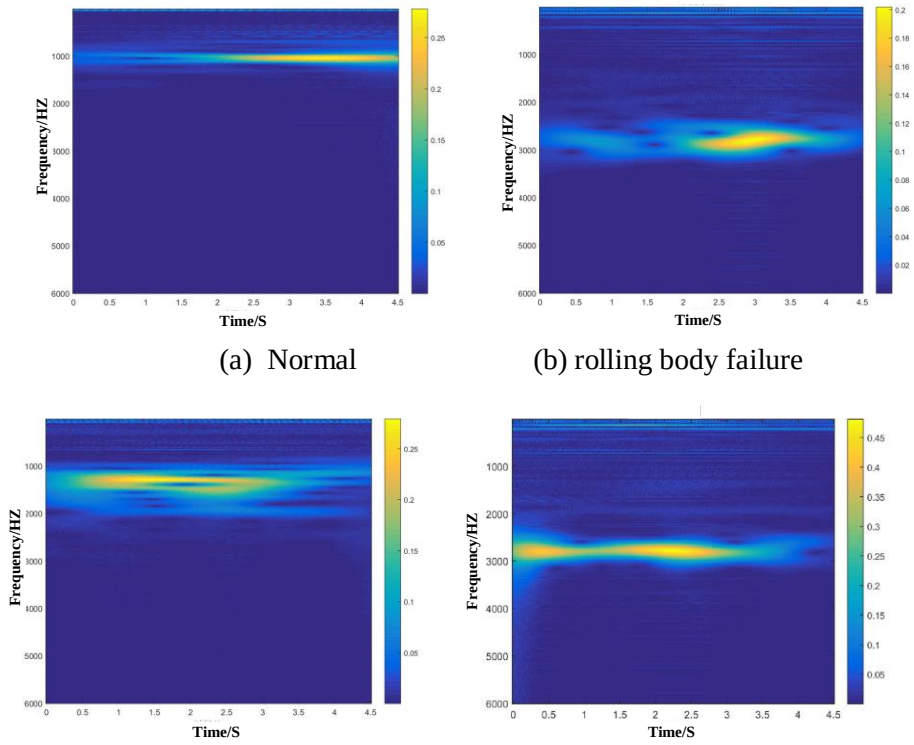
(c)rolling body faults-before (d) rolling body faults-After

Fig 9 CWT time-frequency diagram before and after VMD processing

Fig9 (a) and (b) are the time-frequency diagrams before and after VMD processing under normal conditions, and Fig9 (c) and (d) are the time-frequency diagrams before and after VMD processing in the ball fault state. It is found by comparison that after VMD extracts the fault

features, the fault features of the time-frequency image are more obvious, the time-frequency focus of the time-frequency image is better, the impact is more obvious, and there are more noises before VMD processing, which is not conducive to fault identification.

Then, CWT is used to transform the signal into time-frequency diagram, and the processed time-frequency diagram is shown in Fig10



(a) Normal (b) rolling body failure

(c)Inner Race failure (d) Outer Race@3 failure

Fig 10 The time-frequency diagram of CWT

The processed time-frequency diagram is input into the migrated MobileNet network to train the network. Time-frequency image feature set includes 600 CWT time-frequency images of different parts of faults with a fault diameter of 0.007 inches. The dimension of each image is 224 by 224.The network training parameters were set as follows: batch_size = 16, epoch =20.Under the condition of not adding noise, the rolling bearing fault identification based on vibration signal is realized through MobileNet.

Fig11 shows the error convergence curve of the MobileNet network in the training process, where the abscissor is the number of iterations of all batches of samples, and the ordinate is the recognition rate and loss value. As can be seen from the figure, the loss value of the model tends to 0 and remains stable when the iteration reaches the seventh time. At this point, the model has been trained to the convergence state. As can be seen from the figure, the test set has a recognition rate of 94%.

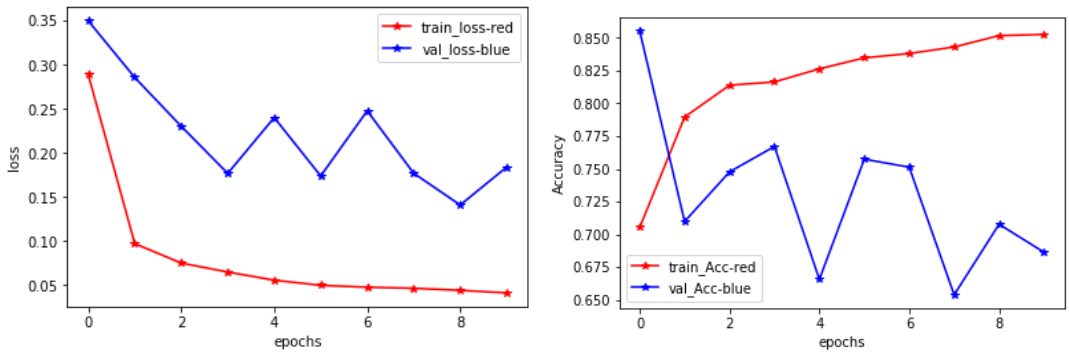


Fig 10 The error convergence curve of VMD-CWT-MobileNet

In order to further prove the effectiveness of the method presented in this paper, time-frequency images without VMD processing are used to train the MobileNet network, and the error convergence curve is shown in Fig11.It can be seen from the figure that the recognition rate of the test set only reaches 68.7%, which proves the superiority of the method in this paper.

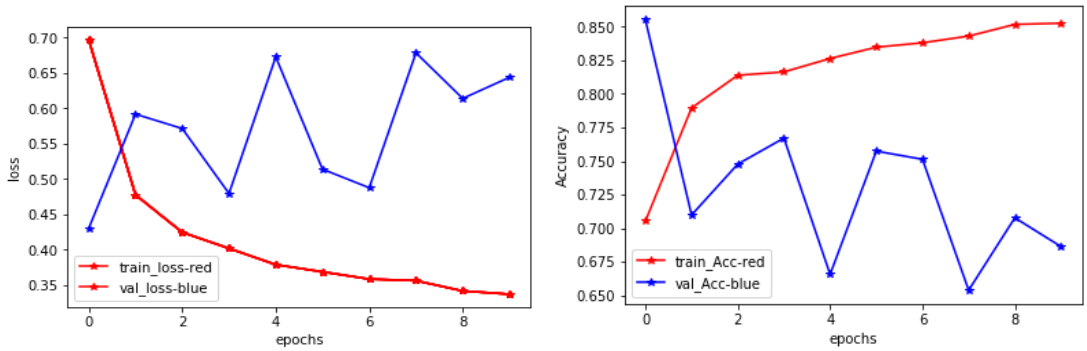


Fig 10 The error convergence curve of MobileNet

5.CONCLUSION

In this paper, the VMD algorithm is used to extract the signal features, and then the wavelet transform is used to extract the time-frequency features. After the data is enhanced, the MobileNet network is trained to accelerate the convergence speed. In this paper, transfer learning is added in the network training process, and the weights of the first several layers which are pretrained on other data sets are transferred to the corresponding network.The following conclusions can be drawn:

(1).After the extraction of fault features by VMD, the fault features of time-frequency images are more obvious, the time-frequency focusing of time-frequency images is better, the impact is more obvious, and there are more noises before VMD processing, which is not conducive to fault identification.

(2).The comparative experiment shows that the fault classification accuracy of the MobileNets network only is 68.7%, and the fault classification accuracy of the proposed method is 94%, which is a great improvement.

(3).The comparison of CNN experiment shows that the training parameters of CNN network are 14,591,685 and the accuracy rate is 95%, while the network training parameters of the method in this paper are 2,264,389 and the accuracy rate is 94%.It is shown that the proposed method can reduce the network training time with a small reduction of accuracy.

REFERENCES

[1] Zhang K , Xu Y , Liao Z , et al. A novel Fast Entrogram and its applications in rolling bearing fault diagnosis[J]. Mechanical Systems and Signal Processing, 2021, 154:107582.

[2] Ma X , Zhou X , An F P . Bi-dimensional empirical mode decomposition (BEMD) and the stopping criterion based on the number and change of extreme points[J]. Journal of ambient intelligence and humanized computing, 2020, 11(2):623-633.

[3] Liu G , Yang C , Liu S , et al. Feature Selection Method Based on Mutual Information and Support Vector Machine[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2021:2150021.

[4] Fault diagnosis of downhole drilling incidents using adaptive observers and statistical change detection[J]. Journal of Process Control, 2015, 30:90-103.

[5] Wang H , Du W . Rolling bearing fault diagnosis based on Slice Energy Entropy Spectral Correlation Density-Continuous Hidden Markov Model[C]// 2019 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC). 2019.

[6] LIU Y, HE B, LIU F, et al. Feature fusion using kernel joint approximate diagonalization of eigen-matrices for rolling bearing fault identification[J]. Journal of Sound and Vibration, 2016, 385: 389-401.

[7] Hoang D T , Kang H J . Rolling element bearing fault diagnosis using convolutional neural network and vibration image[J]. Cognitive Systems Research, 2018, 53(JAN.):42-50.

[8] Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[9] Meghana A S , Sengan S , Arumugam G, Srinivasan P, Kolla Bhanu Prakash. Age and Gender prediction using Convolution, ResNet50 and Inception ResNetV2[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(2):1328-1334.

[10] Tian Y , Liu X . A Deep Adaptive Learning Method for Rolling Bearing Fault Diagnosis Using Immunity[J]. Tsinghua Science and Technology, 2019, 24(006):750-762.

[11] Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[12] Wahab M F , O'Haver T C . Wavelet transforms in separation science for denoising and peak overlap detection[J]. Journal of Separation Science, 2020.

[13] Liu Z , Chai T , Tang J , et al. Signal Analysis of Mill Shell Vibration Based on Variational Modal Decomposition[C]// 2020 39th Chinese Control Conference (CCC). IEEE, 2020.

[14] Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[15] Koonce B . MobileNet v1[M]// Convolutional Neural Networks with Swift for Tensorflow. 2021.

[16] Kumar K K R S , Subramani G , Thangavel S K , et al. A Mobile-Based Framework for Detecting Objects Using SSD-MobileNet in Indoor Environment[M]. 2021.

[17] Saini R , Jha N K , Das B , et al. ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks[C]// IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020.

CLASSIFYING AUTISM SPECTRUM DISORDER USING MACHINE LEARNING MODELS

Tingyan Deng

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA

ABSTRACT

Autistic Spectrum Disorder (ASD) is a very common and serious developmental disability, which impairs the ability to communicate and interact, causing significant social, communication, and behavior challenges. From a rare childhood disorder, ASD has evolved into a disorder that is found, according to the National Institute of Health, in 1% to 2% of the population in high income countries. A potential early and accurate diagnosis can not only help doctors to find the disease early, leading to a more on time treatment to the patient, but also can save significant healthcare costs for the patients. With the rapid growth of ASD cases, many open-source ASD related datasets were created for scientists and doctors to investigate this disease. Autistic Spectrum Disorder Screening Data for Adult is a well-known dataset, which contains 20 features to be utilized for further analysis on the potential cause and prediction of ASD. In this paper, we developed an Autism classification algorithm based on logistic regression model. Our model starts with featuring engineering to extract deep information from the dataset and then applied a modified logistic regression classifier to the data. The model predicts the ASD well in an average F1 score of 0.92.

KEYWORDS

ASD, Classification, Machine Learning, Neurodiversity.

1. INTRODUCTION

Autistic Spectrum Disorder (ASD) is a mental disorder that can affect cognitive, communication and social abilities. Autism is the fastest growing development disability in the world. ASD has been reported to affect as many as 1 in 88 children in the US. Epidemiologic surveys of adult populations suggest that the apparent rise in number of affected children may not represent a true increase in prevalence rates. Nevertheless, there is speculation that broadened definitions, growing awareness, and diagnostic substitution may be contributing to the apparent rise. Regardless of the cause, the current prevalence estimates suggest that there will be more than 2 million individuals in the US with ASD. Up to now, no preventive strategies have demonstrated consistent benefits and no treatments have proven widely efficacious in treating the core symptoms of ASD. Consequently, ASD causes lifelong disabilities for affected individuals and significant burdens on their families, schools, and society.

Raising awareness helps people understand and not be frightened by the disability. Many related researches were investigated to recognize it, prevent it or treat it. In work [1], the authors presented a method using the screening trying to diagnosis this disorder. In work [2], LT Curtis and his partners gave several approaches including nutritional and environmental approaches to prevent the Autism. In work [3], music therapy was used to enable communication and expression and thus can solve some problems of this disorder. In an effort to make easier and

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 59-65, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110306

earlier detection possible, we are using a modified logistic regression model to construct an ASD classification algorithm and the work is discussed in this paper.

1.1. Statement of the Problem

As discussed in the introduction above, the problem is the increasingly ubiquitous occurrence of the ASD symptom, yet little recognitions and efforts were put into solving this issue. With more and more cases of teenager ASDs, we have to come up with a way to ease the pain of ASD. We are aiming to push the barrier of ASD detection knowledge in this study and help future scientists by providing our results.

1.2. Aims/Goals of the Research

In this study, our goal is to provide a faster, and easier machine-learning based approach that can be implemented in future ASD detections and find the related attributes that are causing the ASD. Machine learning has immense potential to enhance diagnostic and intervention research in the behavioral sciences and may be especially useful in investigation involving the highly prevalent and heterogenous syndrome of ASD. In recent years, with more and more advanced computational and engineering methodologies being employed to meet the needs of cross-subject applications, machine learning showed promise in detecting many medical symptoms, which greatly increased the chance of being cured for millions of patients.

Applying the state-of-the-art model is crucial because more and more teenagers have the symptom and there will be a huge amount of middle-age autism community in future and we must detect the symptom earlier so that doctors can cure them earlier.

In our study, we are using a method based on logistic regression model for ASD classification. The specific design of our experiment is listed in the section 2 and section 3.

1.3. Workplan

The work plan, including the experiment/approach, the data we used, and findings are discussed in section 2 and 3.

1.4. Our Contribution and the Flow of the Paper

This paper applies logistic regression to ADS diagnosis. More specifically, the data columns and meanings are described in the first step. Then the data imputation method and feature engineering methods are introduced to further proceed the original data. Data visualization technique was also utilized and the paper contains many easy-to-understand figures. The experiments show the metrics like accuracy, recall, and F1 score. We got an average F1-score of 0.92, which proves the feasibility of our model.

The remainder of this paper is organized as follows. Section II introduces data structure and feature engineering methods. Section III gives a brief introduction to logistic regression models, and the experimental results and analysis. Finally, Section IV gives the summary of whole paper and discuss the ethical concerns of the project.

1.5. Ethics of the Project

Ethics are broadly the set of rules that govern our expectations and of our own and other's behavior. As discussed in [5], research ethnics are important for a numerous of reasons. Firstly, it supports the values required for collaborative work, such as mutual respect and fairness. Secondly, it means that researchers can be held accountable for their actions. Thirdly, good research ethnics support important social and moral values, including the quality of not harming others. If our team members have an internal conflict due to the fact our members have a different opinion towards neurodiversity, I will suggest them to reconcile by taking a look at NISE (Frist Center for Autism and Innovation) website where there are a plethora amount of information about the idea and inspiration behind neurodiversity and also the inspired engineering in this field.

In this project, the dataset we are using is directly downloaded from Kaggle and it's an anonymous, open source dataset, which guarantees the privacy of the research participants. The whole purpose of this project is to accelerate the studies of autism and make the ASD community more inclusive and welcoming. If any ethical dilemmas occur in our project, we will first talk to the people who believes our project violates any computer ethics or ethics in general and then solve the issue by consulting one of the professionals in the NISE community or other experts in the field.

2. DATA AND FEATURE ENGINEERING

The data set can provide a lot of information. In order to explain the dataset intuitively, we list the features of dataset in the table 1.

Table 1. Data Structure

A1_Score to A10_Score
age
gender
ethnicity
jundice
austim
contry_of_res
Used_app_before
result
relation
Class/ASD

As shown in the table 1, the features age, gender, ethnicity are attributes of the ASD testers and easy to figure out the meaning of them. The A1_Score to A10_Score are the answer code of the question based on the screening method used. In particular, they are binary values, either equal to 0 or 1. Statistically, the data has 704 entries and the memory size is 115.6KB.

From the figure 1, we can find the columns ethnicity, relation and age columns have missing values. Since the ratio of the missing values are small, we can drop these missing values or impute the missing values. In this paper, I impute the missing value of age column with the averaged age. And drop missing values of other columns like relation. These steps are necessary for later visualization or building machine learning model.

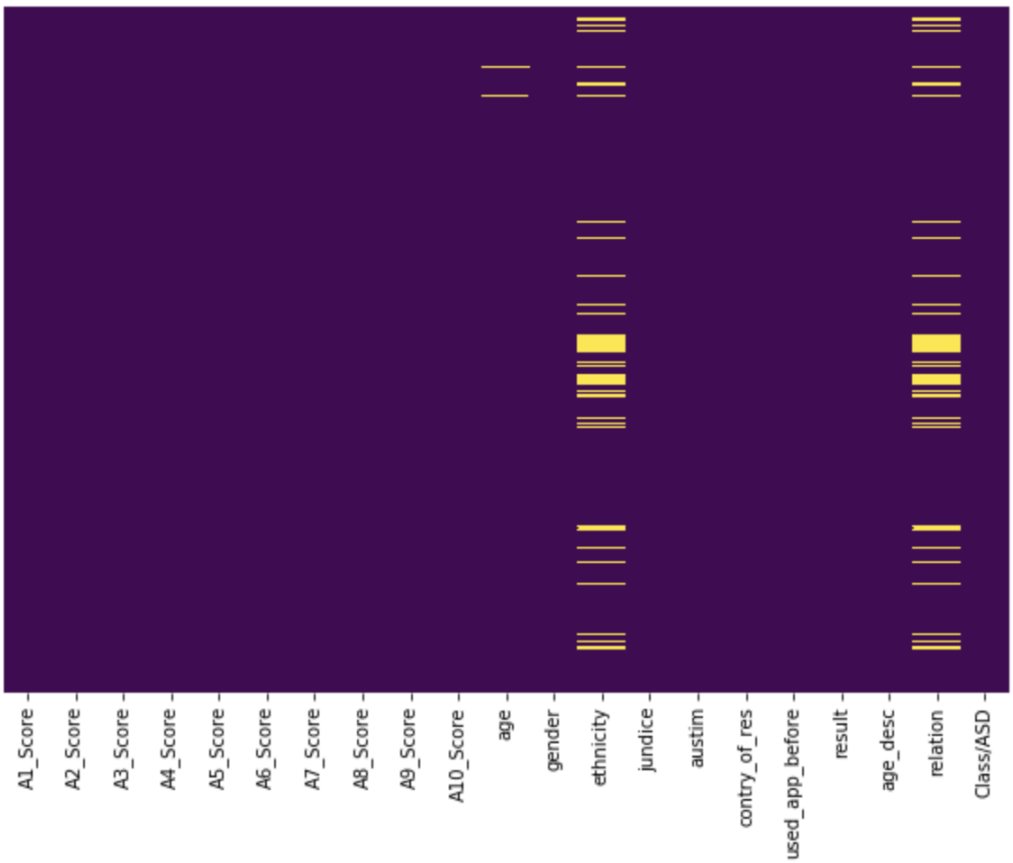


Figure 1. ASD Missing Values Distribution

Besides, we also did some visualization of different features. In the Figure 2, it shows the age distribution. We can find that most of recognized cases are between 0 and 40 years old. This result shows that the ASD mainly occur in young generation other than old people. In figure 3, we can see the ADS Case distribution. The positive sample are higher than negative sample, which means the data is unbalanced data.

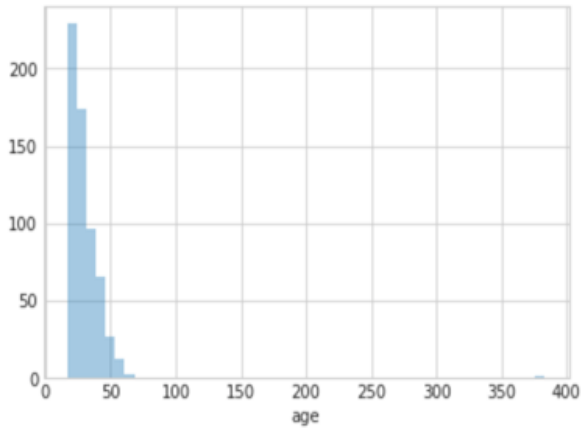


Figure 2. Age Distribution

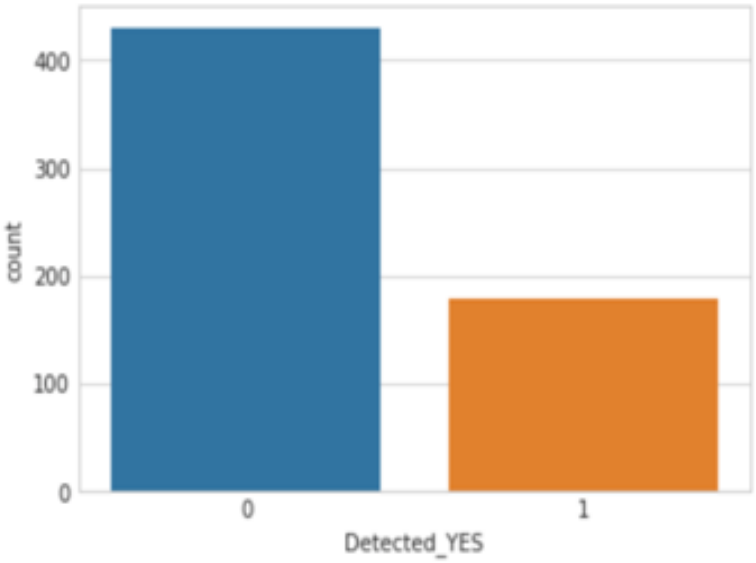


Figure 3. ADS Case Distribution

3. MODELS AND THE EXPERIMENTS

Logistic regression is a linear model which uses a logistic function for classification. The logistic function is as followed:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Equation1. logistic function

The *f(x)* represents the output of the function, L represents the curve's maximum value, k stands for logistic growth rate or steepness of the curve, x0 stands for x value of the sigmoid midpoint, x stands for the real number. Besides, we use the L2 Norm in this model to prevent overfitting.

In the experiment step, we split the data into training set and test set in a ratio of 7:3. Training set is used for fitting the model and test set is used to validate the performance of the model. In the table 2, we list the metrics and performance of our model.

Type	Precision	Recall	F1-score	Count
0	0.98	0.97	0.98	132
1	0.92	0.96	0.94	51
avg	0.97	0.97	0.97	183

Table 2. Performance of our model

As shown in the table 2, the average score of precision is 0.91 and F1-score is around 0.92, which means our model is very accuracy for Autism classification.

4. CONCLUSIONS

In this paper, we propose a method based on logistic regression model for Autism classification. We used the screening data together with meta data like age, gender to fit the model. In Section II, we introduce the data structure and the data size. In addition, the feature engineering is also covered in this part. In Section III, the logistic regression and its function is explained. Then we mentioned the details of the experiments step and the metrics of our model. The experiments show the power of our model since it has a high accuracy. In the future, we are planning on using more machine models like SVM, LightGBM and do a compare and contrast on different models on classification results as discussed in [8].

ACKNOWLEDGEMENTS

We thank the Kaggle platform for provide an ASD screening dataset for researchers to investigate artificial algorithms. Besides, we also appreciate the VUSE (Vanderbilt University School of Engineering) lab for providing the free computation resources like free GPU cards, Rtx 2080 Ti.

REFERENCES

[1] Filipek P A, Accardo P J, Baranek G T, et al. The screening and diagnosis of autistic spectrum disorders[J]. Journal of autism and developmental disorders, 1999, 29(6): 439-484.

[2] Curtis L T, Patel K. Nutritional and environmental approaches to preventing and treating autism and attention deficit hyperactivity disorder (ADHD): a review[J]. The Journal of Alternative and Complementary Medicine, 2008, 14(1): 79-85

[3] Gold C, Wigram T, Elefant C. Music therapy for autistic spectrum disorder[J]. Cochrane Database of Systematic Reviews, 2006 (2).

[4] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

[5] David B. Resnik, J.D., Ph.D, What Is Ethics in Research & Why Is It Important? Retrieved December 10, 2020, from https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

[6] M. F. Misman et al., "Classification of Adults with Autism Spectrum Disorder using Deep Neural Network," 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh, Malaysia, 2019, pp. 29-34, doi: 10.1109/AiDAS47888.2019.8970823.

[7] M. Elbattah, R. Carette, G. Dequen, J. -L. Guérin and F. Cilia, "Learning Clusters in Autism Spectrum Disorder: Image-Based Clustering of Eye-Tracking Scanpaths with Deep Autoencoder," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 1417-1420, doi: 10.1109/EMBC.2019.8856904.

[8] T. Deng, Y. Zhao, S. Wang and H. Yu, "Sales Forecasting Based on LightGBM," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 383-386, doi: 10.1109/ICCECE51280.2021.9342445.

AUTHOR

Tingyan Deng is a junior student at Vanderbilt University studying computer science, mathematics and economics. He is passionate about using technology to make an impact.

BEST PRACTICES IN DESIGNING AND IMPLEMENTING CLOUD AUTHENTICATION SCHEMES

Zhihao Zheng, Yao Zhang, Vinay Gurram,
Jose Salazar Useche, Isabella Roth, Yi Hu

Department of Computer Science, Northern Kentucky University,
Highland Heights, Kentucky USA 41099

ABSTRACT

At present, the development and innovation in any business/engineering field are inseparable from the computer and network infrastructure that supports the core business. The world has been turning into an era of rapid development of information technology. Every year, there are more individuals and companies that start using cloud storages and other cloud services for computing and information storage. Therefore, the security of sensitive information in cloud becomes a very important challenge that needs to be addressed. The cloud authentication is a special form of authentication for today's enterprise IT infrastructure. Cloud applications communicate with the LDAP server which could be an on-premises directory server or an identity management service running on cloud. Due to the complex nature of cloud authentication, an effective and fast authentication scheme is required for successful cloud applications. In this study, we designed several cloud authorization schemes to integrate an on-premises or cloud-based directory service with a cloud application. We also discussed the pros and cons of different approaches to illustrate the best practices on this topic.

KEYWORDS

Cloud Application Authentication, Identity Management in Cloud, IAM.

1. INTRODUCTION

With the development of science and technology, more and more people are now using cloud services, such as cloud storage, cloud database, etc. The concept of cloud computing has been applied in a lot of fields, such as finance, medicine, education, and manufacture [3]. Cloud computing - at least as an extension of virtualization, has become increasingly influential. The advantages in cloud technology are that large amount of data can be stored and computed at a very low cost. Cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) can be rapidly provisioned and released with minimal management effort or service provider interaction [1].

Identity management in the cloud is particularly difficult since the identity itself has cross-border features. At the same time, the identity management could have a critical impact from both architectural and organizational point of view. Many businesses are afraid of using the cloud because it would expose themselves to possible attacks and data corruption [2]. In addition, many companies do not have sufficient resources to manage identity authentication in the cloud because they lack flexible identity management to cover both on-premises and cloud native

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 67-76, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110307

applications. Due to the complex nature of cloud authentication, an effective and fast authentication scheme is required for a successful cloud application [5]. When implementing a cloud authentication scheme to integrate an on-premises or cloud-based directory service with a cloud application, we aimed to discuss the pros and cons of different approaches to illustrate the best practices on this topic.

2. BACKGROUND AND MOTIVATIONS

2.1. Standards to Improve Scalability

Cloud computing allows people to access a configurable pool of computing resources—networks, servers, storage, applications, and services over the network at any time and on a convenient, ready-to-use basis. These can be quickly prepared and published with minimal administrative effort, or interaction with service providers [2]. It allows organizations to instantly add processing power or functionality while not investing in new infrastructure, training new employees, or purchasing new software licenses [4].

Cloud computing covers any subscription-based service that extends existing IT capabilities on the Internet in real time. Public cloud usually refers to software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS). A private cloud is an application or platform that is specific to a particular organization and deployed on its own site, often hidden behind a firewall [3].

2.2. Performance Scalability

When we think about scalability, most people immediately think of how a system handles large-scale transactions, floating-point operations per second, and so on. A key feature of the cloud is its ability to meet changing needs by adding or subtracting computing power, providing elastic scalability [7]. For example, when it is urgent to require large-scale computing power, an application can be upgraded quickly on the Amazon Elastic Compute Cloud (EC2) to use a virtual machine with larger computing and storage capabilities. To extend this concept a little further, the new cloud application is designed to support the linear expansion of the architecture of N+1, which can support nearly infinite scale of computing operations.

2.3. Integrate and Manage Scalability

A not-so-comprehensive and scale-related challenge is about how quickly organizations can deploy, integrate, and manage a system over time [6]. When the system causes friction - such as in administrative tasks, this can cause system scalability barriers, especially for identity management [4]. If the infrastructure is defined as common hardware, software, and network services that generate IT capabilities within the enterprise, the identity management infrastructure includes directory services, identity and access management services, network proxies, and verification systems used throughout the enterprise. Many companies today are trying hard to build an identity infrastructure that can work under a cloud architecture and can gracefully upgrade in a cloud fashion.

To be able to upgrade to meet cloud architecture and growth needs, system architects must focus on optimal management and integration of identities [2]. Identity management is a key bottleneck for adopting the cloud for many businesses. Architects understand that their vision must go beyond the basic performance level of cloud scalability, and also design a strategy that allows the management and integration of identities to be scalable.

2.4. A Cloud-level Identity Structure

Through different technologies, standards, and use cases, the identity verification can be obtained across the previously separately managed security domains [8]. So that users in one domain can securely and seamlessly access data and systems in another domain without the need for redundant user management. With this federated identity, many elements and fields are intertwined, just like weaving cloths.

In the past, organizations stored network identities in various directories and identity databases. With the growth of the Internet and the emergence of cloud applications, they have discovered that they need to manage identities outside the traditional network [6]. Today's network administrators must manage multiple accounts for corporate and cloud applications. This duplication of labor increases the workload and leads to security risks due to administrators having to manage multiple user identities and passwords. Cooperation with outside partners and contractors also requires the company to open up network boundaries to outsiders.

To ensure the security of so many information assets and data, companies need to seamlessly use identity management to connect to the cloud. To achieve successful cloud identity management, the industry must ensure that the identity meets the unique architectural needs of the cloud, and identities are regarded as a structure that will be integrated, abstracted, and extended. The identity is delivered as SaaS, just like the cloud platform itself supports [5].

3. OUR MODEL AND APPROACHES

3.1. An Abstract Concept

To implement a cloud-level identity structure, it requires the abstraction of identities into identity services [9]. Application developers historically plug identities into the application itself and maintain a local user base to perform authentication. This leads to redundant and often stale data, passwords, and greater help center overhead.

In the past decade, applications began to externalize identity management, starting with an external directory that intensively authenticates users based on Lightweight Directory Access Protocol (LDAP). This is an important step for the scalability of identity management, but we need to do more - LDAP password authentication is not enough. Businesses must be able to use more than one type of certification, depending on the level of threats to be applied [7].

3.2. Problem Description

With the increasing of information processing and storage needs, enterprise users have more and more demands for the efficient information synchronization and service collaborations. Nowadays, cloud storage and cloud computing has become a popular resource. The cloud has the characteristics of convenient resource sharing, low maintenance and management cost, and large scale. Enterprises are more likely starting to build their cloud data center. However, most of the enterprise data are still stored locally and cannot be perfectly connected with public or private clouds [10]. How to integrate local and cloud storage easily and improve the utilization rate of cloud resources is a problem that many enterprises are facing. Most companies do not store all data on the cloud, since some of company’s data are highly classified like bank’s credit card information. So organizations will keep some important data on the local server, and other data will be saved on cloud storage.

In our experience, we are going to set up three different solutions to test which solution could provide us high performance, high security and low cost. The tools we are going to use are Windows sever 2016 and Amazon Web Services (AWS). Windows sever 2016 is a server operating system, and Active Directory Federation Services is possible to configure AD FS to authenticate users stored in non-AD directories. AWS (Amazon Web Services) provide services to individuals, companies and governments [8].

Storing data locally means the enterprise has its own dedicated data center. Traditionally, this is how most organizations design and maintain networks. Regardless of the other aspects, this requires physical hardware, the space required for the hardware, and backup and disaster recovery services [12]. On the another hand, some enterprises choose to store data in the cloud. Cloud is actually a server network that serves different functions from each server. Some servers store data and some run applications. We could easily notice that we are tending to not buy boxed software from stores, but we pay the monthly fee on the Internet access platform. This is a real running cloud.

In this study, we designed cloud authorization schemes to integrate an on-premises or cloud-based directory service with a cloud application. We also discussed the pros and cons of different approaches to illustrate the best practices on this topic.

The Amazon Web Services (AWS) is a professional cloud computing service that offered by Amazon [11]. It was launched in 2006 to provide IT infrastructure Services to businesses in the form of Web Services. The services that AWS provides including elastic compute cloud (Amazon EC2), Simple Storage Service (Amazon S3), Simple database (Amazon backs), Simple Queue Service and the Amazon CloudFront… etc. As the largest services provider, Amazon AWS provides infrastructure and services that build reliable, fault-tolerant, high-availability systems in the cloud [13].

The cloud application communicates with the LDAP server which could be an on-premises directory server or an identity management service running on cloud. Due to the complex nature of cloud authentication, an effective and fast authentication scheme is required for a successful cloud application.

3.3. Models and Procedures

3.3.1. System Architecture

The foundation of the system involves setting up the Windows Server 2016 and building sample Rails login web application by using AWS for Rails Developers. This step is basically for us to get used with everything we were going to use such as Windows Server 2016, Amazon web service (AWS), and Ruby.

We used the SDK with Ruby on Rails. The AWS SDK for Ruby helps take the complexity out of coding by providing Ruby classes for almost all AWS services, including Amazon Simple Storage Service, Amazon Elastic Compute Cloud, and Amazon DynamoDB [9]. Before we use it, we need to install the AWS SDK for Ruby. Then we tried a couple of commands to test whether it works correctly. Such as creating a bucket, adding files to the bucket, listing the contents of that bucket, etc.

Then we configured the AWS SDK for Ruby by using the AWS access keys. We also set up the AWS credentials. After that we set up the region and nonstandard endpoint.

We integrated the AWS SDK for Ruby with Rails, then we used the Amazon SES that support for ActionMailer. After that we could log in by entering the target webpage on the browser. The screenshot illustrated in Figure 1 shows the login page to be adapted with different authentication schemes.

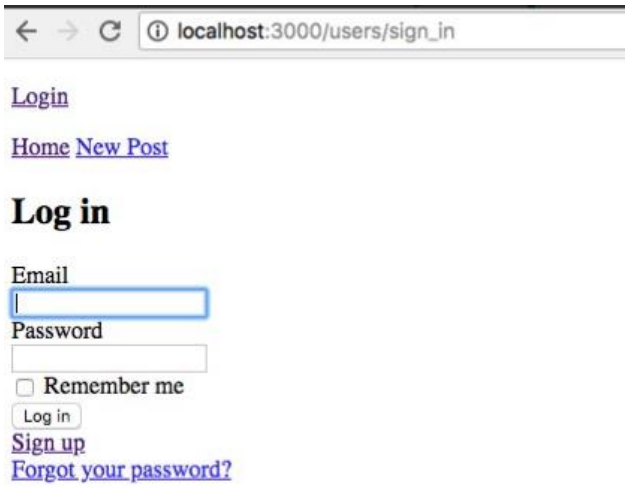


Figure 1. Login Page to be Adapted with Different Authentication Schemes

3.3.2. Model 1: On-Premises Authentication Service with a Cloud Application

Then, we set up the first solution: on-premises authentication service with a cloud application: we used ADFS connected with AWS, enabling federation to AWS using Windows Active Directory, ADFS, and SAML 2.0.

For this step, we used the active directory federation service (ADFS) to make connection with the cloud application. We used the Amazon Elastic Beanstalk that allows users to quickly deploy and manage their applications without configuring the infrastructure that runs those applications [11]. The connection between authentication service and the application is shown in Figure 2.

Firstly, we installed the AD Federation Service on Windows Server 2016. Then we enabled our users to access Office 365 with AWS managed Microsoft AD. We added two containers by ADFS to the AWS Microsoft AD. Then we installed the ADFS, we integrated ADFS with AD [8].

After that, we deployed a Ruby on Rails application to Elastic Beanstalk [9]. These are the steps we did:

1. Create a Rails App to Deploy
2. Create an Application on Elastic Beanstalk
3. Install AWS CLI and EB CLI
4. Create an Environment on Elastic Beanstalk
5. Set Up an RDS Database
6. Observe Our Working App

Although there are too many choices to pick from for this step, it’s very important to use the Rails application directory when running the AWS and eb commands.

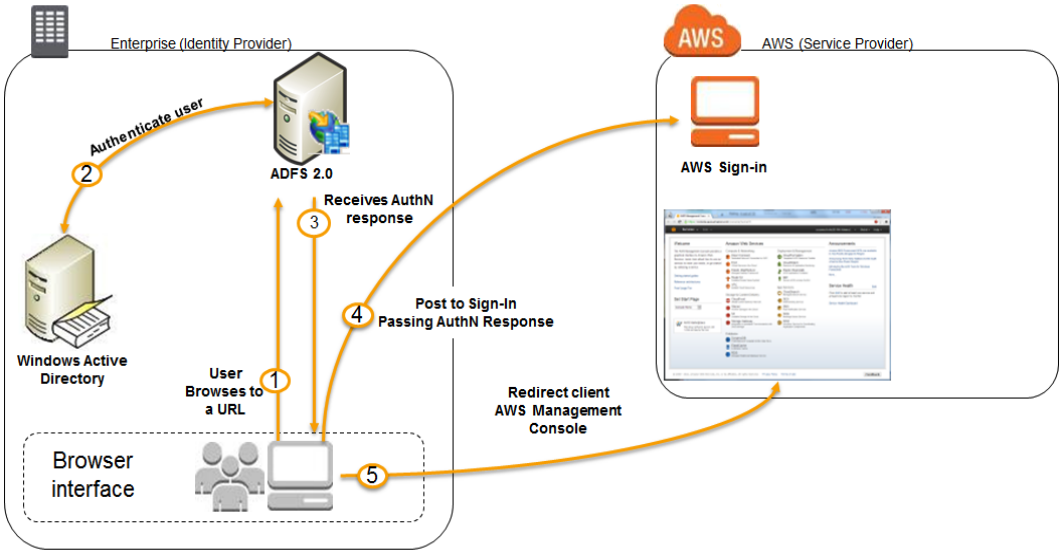


Figure 2. On-Premises Authentication Service with a Cloud Application [14]

3.3.3. Model 2: Third-Party Authentication Service with a Cloud Application

Next, we set up the second solution: third-party authentication service with a cloud application: build and deploy a federated internet identity application with AWS Elastic Beanstalk and then log into it with Amazon.

For this step, the third-party service we used is the Amazon account. We were trying to let our users use their Amazon account to log in when they were on our website. Since Amazon is very popular and people almost using Amazon every single day, it is easy to connect with Amazon instead of other account such as Google or Facebook.

To build and deploy a federated web identity application with AWS Elastic Beanstalk and Login with Amazon [12], we did the following steps as shown in Figure 3. This is the identity authentication scheme in the authentication process, the following picture illustrates the entire process.

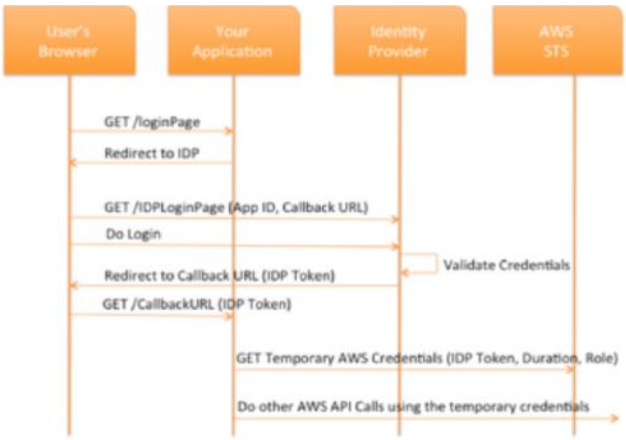


Figure 3. Identity Federation [15]

To develop a web-based application using Federated Web Identity, we firstly deploy the application to the Amazon Elastic Beanstalk. Secondly, we register our application with the Amazon Identity Provider. Then for our application, we need to define the permission in AWS. After that, for the load balancer, we configured SSL certificate. Then we configured our application on AWS. In the end, we tested our application to get the login page adapted with the third-party authentication service as shown in Figure 4.

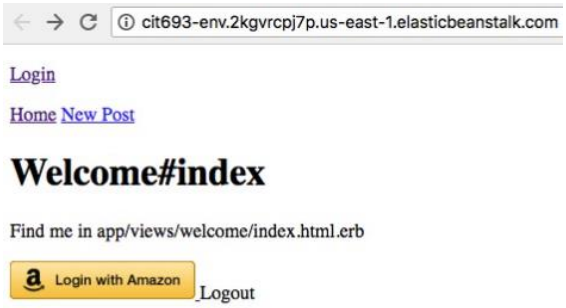


Figure 4. Login Page Adapted with Third-Party Authentication Service

3.3.4. Model 3: Cloud Authentication Service with a Cloud Application

In this step, we set up the third solution: cloud authentication service with a cloud application: log on to AWS services by using on-premises active directory: For this step, we used the AWS Directory Service for Microsoft Active Directory, it also known as AWS Microsoft AD. It enables our directory-aware workloads and AWS resources to use managed Active Directory in the AWS Cloud [13]. AWS Microsoft AD is built on actual Microsoft Active Directory and does not require us to synchronize or replicate data from your existing Active Directory to the cloud [11]. Figure 5 and 6 shows the AWS directory service for Active Directory and cloud directory details.

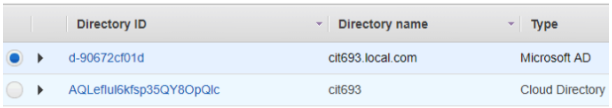


Figure 5. Directory Names and IDs for AWS Directory Service for Active Directory

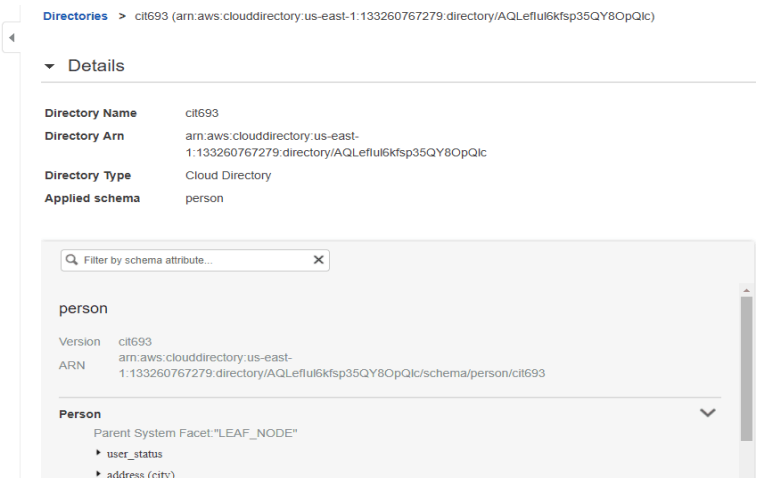


Figure 6. AWS Cloud Directory Example

We also created an Angular2 web application and implemented a third party authentication method. With the purpose of implementing a third party application to achieve authentication, we created a simple Angular2 Single-Paged web application.

Taking advantage of the fact that Angular2 was conceptualized to work as single components for better performance and simplicity in writing code, we created three components. The first component, containing the "Home" view, which only renders a simple line of text.

The second component, called the "Navbar," renders a navigation bar where some buttons will be placed that when pressed will call a service which will display a series of options for the user to decide which profile and/or social network will be used to later perform authentication. The third component, called "protected," will render information retrieved from the profile used to perform authentication once the user is properly authenticated.

In order to successfully perform authentication, a third party application called "Auth0" was implemented. Auth0 offers a free subscription as long as there are less than 7000 active users. Auth0 is not only very easy to be integrated as an authentication service, but also serves as an universal authentication layer for both on-premises and cloud native applications.

4. RESULTS AND CASE ANALYSIS

The use of these three cloud authentication solutions is highly dependent on the size of the institution/organization. Table 1 illustrates the comparison result based on cost, security, ease of maintenance, and client data management.

Table 1. Cloud Authentication Comparison

	Cost	Security	Maintenance	Client Data Management
ADFS	High	Highly Secure	Hard	Hard to management, but easy to obtain data
Third-Party	Median	Relatively not Secure	Easy	Easy to management, but hard to obtain data
AWS Cloud Directory	Low	Secure	Relatively Easy or Median	Easy to management, but hard to obtain data

We have some cloud authentication suggestions for these three kinds of clients below:

1. For start-up companies: we should mostly think about the cost and the number of customers. So, using the third-party and AWS directory solution not only can reduce the cost, but also give you a chance to attract customers from some third-party platforms. In addition, the maintenance will be relatively easy since they don't have to hire too many employees to manage these account and data. For the maintenance of server, they don't have to hire people as well since the third-party and AWS directory service have already cover the maintenance as well.

2. For median-size companies: for this kind of companies/organizations, they may already have the local active directory. So that they can easily use the ADFS to authenticate the

web applications. For example, some organizations like university will highly likely to use this solution. On another hand, for those companies that are doing business/entertainment, they may want to attract clients from third-party web applications. they can easily add the third-party connection on their web apps.

3. For large organizations: in this case, they usually have a large group of cloud service professionals, great budget, as well as multi-cloud services. They are very likely to use federated authentication schemes. We recommend them to add the third-party solution to make it more convenience for their clients to login.

5. CONCLUSIONS

In this study, we set up three different cloud authentication solutions to test which solution could provide us high performance, high security and low cost. We designed several cloud authorization schemes to integrate an on-premises or cloud-based directory service with a cloud application. We also discussed the pros and cons of different approaches to illustrate the best practices on this topic. Our goal is to illustrate the best practice on cloud authentication based on usage scenarios.

ACKNOWLEDGEMENTS

We greatly appreciate Dr. Traian Marius Truta, for helping review the draft version of the paper and providing comments.

REFERENCES

[1] D. Chang, M. Benantar, J. Chang, V. Venkataramappa, “Authentication and authorization methods for cloud computing security”, USPTO Patent Grants, July 2014.
[2] D. Chadwick, K. Fatema, “A privacy preserving authorisation system for the cloud”, Journal Of Computer And System Sciences, Canterbury, United Kingdom, 2013.
[3] W. Smari, A. Navarro, W. McQuay, B. Tang, R. Sandhu, Q. Li, “Multi-tenancy authorization models for collaborative cloud services”, Concurrency and Computation, November 2014.
[4] E. Bertino, F. Paci, R. Ferrini, and N. Shang, “Privacy-Preserving Digital Identity Management for Cloud Computing”, In IEEE Data Engineering, pages 21–27, 2009.
[5] M. Dragos Marian, “Cloud Identity and Access Management – A Model Proposal.”, Journal Of Accounting And Management Information Systems, Romania, 2012
[6] T. Piepers, “Cloud Identity & Access Management Model: success factors for Identity & Access Management in cloud computing.”, Masters Thesis, Netherlands, 2013.
[7] S. Smita and M. Deep, "Identity Management issues in Cloud Computing", International Journal of Computer Trends and Technology (IJCTT), vol. 9, no. 8, 2014.
[8] L. Song, C. Jie, Z. Hong, L. Lu, “Public Auditing with Privacy Protection in a Multi-User Model of Cloud-Assisted Body Sensor Networks”, Sensors, Vol. 17 Issue 5, p1-19, Ipswich, MA, May 2017.
[9] F. Nzanywayingoma, Y. Yang, “Efficient Resource Management techniques in Cloud Computing Environment: Review and discussion.”, Telkomnika, Vol. 15 No. 4, pp. 1917-1933, Beijing, China, December 2017.
[10] S. Rizwana, M. Nerul Navi, M. S, M. Kharghar Navi, “Identity Management in Cloud Computing”, International Journal of Computer Applications, Vol. 63, No. 11, 2013.
[11] U. Habiba, R. Masood, M. Shibli, M. Niazi, “Cloud identity management security issues & solutions: a taxonomy”, Complex Adapt Syst Model, Vol 2, No. 5, 2014.
[12] M. Darwish, A. Ouda, L. Capretz, “Cloud-Based Secure Authentication (CSA) Protocol Suite for Defense against DoS Attacks.”, Journal of Information Security and Applications, 2015.
[13] P. Cigoj, B. Blažič, “An Authentication and Authorization Solution for a Multiplatform Cloud Environment. Information Security Journal: A Global Perspective”, vol. 24, no. 4-6, pp. 146-156, Ljubljana, Slovenia, August 2015.

[14] AWS, Enabling Federation to AWS Using Windows Active Directory, ADFS, and SAML 2.0. https://aws.amazon.com/blogs/security/enabling-federation-to-aws-using-windows-active-directory-adfs-and-saml-2-0/
[15] AWS, Identity federation. https://aws.amazon.com/identity/federation/

AUTHORS

Zhihao Zheng was a graduate student at Northern Kentucky University majoring in computer information technology. He is interested in Database Systems, Data Security, Data Mining. He has worked on several projects such as AWS solution Architecture, Data Mining in Python and Weka, Effect of Virtualization on System Performance, etc.

Yao Zhang was a graduate student at Northern Kentucky University majoring in computer information technology. She is interested in Database Systems, Data Security, Data Mining. She has worked on several projects such as AWS solution Architecture, Design and Analysis of Experiments, Applied Mathematical Models, etc.

Jose Salazar Useche is an undergraduate student at Northern Kentucky University majoring in Computer Science. He is interested in Machine Learning, Computer Vision and Quantum Computing. He has worked on several projects such as a Yoga postures classifier using Google's AIY Vision kit, a Spotify-like music catalog and a task manager app.

Dr. Yi Hu is a Professor of Computer Science at Northern Kentucky University. He is also a CISSP and CEH. He has published more than 30 papers on security and trust management. In addition, he is the Director of Center for Information Security at NKU and Director of Kentucky Collegiate Cyber Defense Competition with extensive experience on promoting security education and awareness.

Vinay Gurram was a graduate student at Northern Kentucky University. His research interests are cloud computing and cybersecurity.

Isabella Roth is an undergraduate student at Northern Kentucky University. Her research interests are cloud authentication and cloud data security.

INVESTIGATING DATA SHARING IN SPEECH RECOGNITION FOR AN UNDER-RESOURCED LANGUAGE: THE CASE OF ALGERIAN DIALECT

Mohamed Amine Menacer and Kamel Smaïli

Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

ABSTRACT

The Arabic language has many varieties, including its standard form, Modern Standard Arabic (MSA), and its spoken forms, namely the dialects. Those dialects are representative examples of under-resourced languages for which automatic speech recognition is considered as an unresolved issue. To address this issue, we recorded several hours of spoken Algerian dialect and used them to train a baseline model. This model was boosted afterwards by taking advantage of other languages that impact this dialect by integrating their data in one large corpus and by investigating three approaches: multilingual training, multitask learning and transfer learning. The best performance was achieved using a limited and balanced amount of acoustic data from each additional language, as compared to the data size of the studied dialect. This approach led to an improvement of 3.8% in terms of word error rate in comparison to the baseline system trained only on the dialect data.

KEYWORDS

Automatic speech recognition, Algerian dialect, MSA, multilingual training, multitask learning, transfer learning.

1. INTRODUCTION

Arabic language comprises thirty modern varieties[1], including its standard form, Modern Standard Arabic (MSA), which is derived from Classical Arabic. MSA is a simplified version of the Classical Arabic (a literary form) with a modernized vocabulary. It is the official form used in the newspapers and in the formal communications. The other Arabic language varieties, referred as dialects, come from historical interactions between classical Arabic and languages of the regional cultures and from the linguistic influence due to colonization. They are used in the Arab world in informal conversational context and in the daily communication.

In many Natural Language Processing (NLP) applications, the bulk of works proposed in the literature is intended for MSA, less works are dedicated to Arabic dialects. For a long time, the NLP community was not interested by Arabic dialects, but nowadays a craze for these dialects has been observed. In fact, there are several reasons for that: the Arabic dialects constitute the daily language of communication in Arab world, they are under-resourced languages, there is no standardization for writing them, some of them are very different from MSA, they often are code-switched, etc. All these features make them challenging in point of view of NLP. In this article,

[1] Source: ISO 639-3 ara documentation

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 77-89, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110308

we focus on an Algerian Arabic dialect, the one used in Algiers and its periphery, for which we propose an Automatic Speech Recognition (ASR) system.

NLP for under-resourced languages, such as Arabic dialects, requires more sophisticated techniques that go far beyond the basic re-training of the models dedicated to well-resourced languages. The approaches that have been proposed so far to recognize under-resourced languages focused mainly on two aspects: proposing, on one hand, data collection methodologies and introduce, on the other hand, advanced training techniques to cope with the lack of data. To develop an ASR system for an under-resourced language, one needs firstly to collect the necessary data for its different components. Works on data collection are carried out via crowdsourcing [1] or via exploring data for which information are shared between languages[2, 3].

For the acoustic data, it is often difficult to obtain spoken transcribed resources for under-resourced languages. One can achieve this by transcribing existing audio resources [4] or by recording speech from existing textual data [5]. Concerning textual data, the easy way to collect them is to investigate web content [6].

Moreover, a pronunciation dictionary must be created; the grapheme-based approach is the simple way to produce it. One considers for Arabic that the pronunciation of each word is simply its grapheme decomposition, and therefore, graphemes represent the basic units for the Acoustic Model (AM) [3, 7]. Other approaches are used to convert graphemes to phonemes such as those based on statistical machine translation [8, 9] or on linguistic rules[10, 11].

Since the data collection for under-resourced languages is time consuming, unsupervised or semi-supervised approaches are pretty adequate in this context. One underlying technique that can be used when only a small amount of transcribed data is available is to develop a baseline ASR system and use afterwards this system to transcribe a large amount of data. These new transcribed data can be used to fine tune the baseline system and improve the speech recognition performance [12]. Another interesting approach is to take advantage from other languages. The idea is to develop a multilingual model that combine information from several languages that share words [2, 13].

For the Algerian dialect, there is no transcribed data for training the acoustic model. To handle this issue, we propose to record a small spoken corpus for developing a baseline system and then, to improve it by taking advantages from the speech data of other languages that impact the Algerian dialect.

2. ISSUES FOR DEVELOPING AN ASR SYSTEM FOR AN ALGERIAN DIALECT

The vocabulary used in the Algerian dialect comes from the historical interaction between multiple languages, namely MSA, French, Turkish and Berber. Words from these languages could be employed without any modification, or they could be altered to produce new words. This fact leads to a new language variety that is different from the MSA, and that can be defined as a mixture of several languages.

Because of the borrowed words, the phonetic system of the Algerian dialect is a mixture of Arabic phonemes and others especially used in the French language. This leads to an exhaustive list of 47 phonemes (34 Arabic phonemes plus 13 French phonemes). Consequently, to correctly recognize the Algerian dialect, the first issue that we need to handle is to train an acoustic model that recognizes all these phonemes.

The Algerian Arabic dialect is mainly spoken, that means that the way of writing is free. People could use Latin or Arabic script or mix all the foregoing in the same sentence to convey their ideas. Some examples of the writing system extracted from social networks are illustrated inTable 1.

Table 1.Examples of some writhing possibilities in the Algerian dialect.

Arabic script	الله يخليك عندي مشكل في ترتيب لفاليز دياولي كاش فكرة
Latin script	Alla hyekhalik aandi mochkil fitartib les valises dyawli kache fekra
Mix script	الله يخليك عندي problème في ترتيب les valisesديياولي كاش idée
Translation	Please, I have a problem of organizing my suitcases, any idea!

In the following sections, the techniques used to model the acoustic and the language aspects for the Algerian dialect are set forth.

3. LANGUAGE MODELLING

Algerian dialect is mainly spoken without any conventional writing rules. Consequently, it is difficult to find well-formed text written in dialect. One way to deal with this issue is to retrieve textual data from social networks. In our previous works, two corpora containing Algerian dialects were constituted: PADIC [14, 15] and CALYOU [6] corpora.

- **PADIC** is a collection of 6400 Modern Arabic sentences with their translations in several Arabic dialects (Two from Algeria, Tunisian, Moroccan, Palestinian and Syrian). This corpus was developed manually by translating Arabic conversational sentences into the different dialect variants.
- **CALYOU** is a large corpus collected from comments of Algerian videos on YouTube. It contains 1.4M dialect sentences written in Arabic and Latin scripts.

While the writing system in PADIC corpus is standardized (by adopting some rules and by using Arabic characters extended with (پ/p/, ڤ/v/, ڤ/g/) for non-letters sounds), sentences in CALYOU corpus are not normalized, since it is a collection of comments extracted from social network, where the way of writing is free. For this reason, we carried out a pre-processing to normalize the data of CALYOU, it consists of:

- Removing all the sentences written or containing Latin script.
- All the homophones that have the same meaning are replaced by the most frequent spelling by using a lexicon proposed in[16]. Some examples are given in Table 2.

After having processed the CALYOU corpus, the total number of sentences is reduced to 650K.

Table 2. Examples of some homophones that have the same meaning.

Homophones	Replaced by	Translation
فيلم – فليم - فلم	فيلم	Film
منفقين - منافقين	منافقين	Hypocrites
خااوة – خوة – خوا – خاوى – خاوة – خاوا	خاوة	Brothers

The training of the Language Model (LM) for the Algerian dialect is not restricted on the two corpora PADIC and CALYOU, we also take advantage from MSA data. Since the amount of the different textual data is unbalanced, the LM, we propose, is a linear combination of four bigram models. Two of them are trained on MSA textual data: the MSA version of Gigaword (1 billion

of word occurrences) and the transcripts of the MSA speech data used to train the acoustic models (315k words). The two others are trained on dialectal data: PADIC and CALYOU. The weights of the linear interpolation are estimated on a development corpus composed by a mixture of MSA and dialect data. The resulting weights for each corpus are 0.48 for CALYOU, 0.22 for MSA Gigaword, 0.11 for PADIC and 0.19 for the transcripts of the MSA speech data.

4. PRONUNCIATION MODELLING

The lexicon is composed by the union of the most frequent words extracted from each dataset used for training the language model. For each word in the lexicon, one needs to have all its pronunciation variants. The issue is how to produce all possible pronunciation variants for Arabic words, among them a subset of dialect words, knowing that Arabic texts are written without any diacritic.

Since linguistic resources are available for MSA, we used an external lexicon [17] as a lookup table from which the pronunciations of the MSA words are extracted and inserted into the pronunciation lexicon of our ASR system. Unfortunately, we do not have the equivalent for the Algerian dialect. For this, we adopted a G2P approach to produce pronunciation variants for dialectal words. We adapted to our purpose the approach proposed in [10]. The conversion G2P process is based on two stages:

- Restore diacritics using a statistical approach. This issue is considered as a machine translation problem where the source language is a set of undiacritized texts and the target one is a set of diacritized texts. A Statistical Machine Translation (SMT) system was trained by using existing tools on a parallel corpus of undiacritized and diacritized Algerian dialect texts. Since this parallel corpus was built manually and the task of vocalization is time consuming, this corpus contains only 4k sentences. This approach led to a precision of 98% at the character level and 96% at the word level.
- Use a set of hand-crafted rules to produce the phonetic representation of the dialectal words. For further details about these rules, the reader is directed to the work [10].

5. ACOUSTIC MODELLING

The main challenge that we are facing is to get a spoken transcribed corpus for the Algerian dialect. Because recording is a costly task, we selected only 4.6k dialect sentences extracted from PADIC and CALYOU and we asked native Algerian speakers to record this small corpus. The selection is carried out in such a way that the length of the sentences fluctuates between 3 and 20 words with an average duration of 4.5 seconds.

Seven Algerian native speakers recorded, in a quiet room and using a professional unidirectional microphone, the selected corpus. Two of them are female and five are male.

The resulted corpus contains 6 hours of speech sampled at 16 kHz. This dataset, named ADIC (Algerian DIalect Corpus) is split into three parts as it is shown in Table 3. The speakers of the Test data are different from the ones of the Train and the Dev data.

Table 3. Some characteristics of ADIC.

Subset	Duration	Speakers
Train	240 min	4
Dev	40 min	
Test	75 min	3

5.1. Learning by using a TDNN architecture

We propose to use an acoustic model based on the time delay neural network (TDNN) architecture [18] as described in Table 4. TDNN is a kind of feed-forward neural network used to better handle the context information of speech signal through a carefully designed hierarchical structure [19]. It is based on the use of context windows where the input layer processes acoustic features with narrow contexts while wider contexts are processed by the deeper layers.

Each deep layer receives several outputs that are spliced from the previous layer. The first layer receives a concatenation of 5 acoustic features corresponding to the features from $t-2$ to $t+2$ (see line 2 of Table 4). In layer 2, we splice together the input at the current frame minus 1 until the current frame plus 2. This means that the second layer will capture implicitly a larger context of the acoustic features from $t-3$to $t+4$.

In this case, one can understand that the number of parameters to train is huge. To deal with this issue, we adopt the method proposed in [18] to sub-sample the TDNN network. In this approach, instead of splicing all the frames, only two frames are gathered corresponding to the first and the last frame of the original method. For instance, the notation $\{-1,+2\}$ in the third line of Table 4, means that only the two outputs -1 and $+2$ are spliced. At the end, the output of the last layer handles implicitly the context of $[t-16, t+11]$ for each acoustic parameter at t timestamp.

Table 4.Context specification for each layer of the TDNN model.

Layer	**1**	**2**	**3**	**4**	**5**	**6**
Input context	$[-2,+2]$	$[-1,+2]$	$[-3,+3]$	$[-3,+2]$	$[-7,+2]$	$\{0\}$
Input context with sub-sampling	$\{-2,+2\}$	$\{-1,+2\}$	$\{-3,+3\}$	$\{-3,+2\}$	$\{-7,+2\}$	$\{0\}$

The training of the TDNN model is based on sMBR sequence-discriminative criterion [20] and the parameters are estimated with the stochastic gradient descent algorithm.

Since ADIC is considered small for training the TDNN model, our idea is to benefit of other languages that impact the dialect (MSA and French) and to transfer the acquired knowledge to the ASR of the dialect. To do so, we proposed three different approaches depending on how the MSA and French acoustic data are integrated into the training process of the acoustic model of the Algerian dialect. These approaches are the multilingual training, the multitask learning and the transfer learning (see Figure 1).

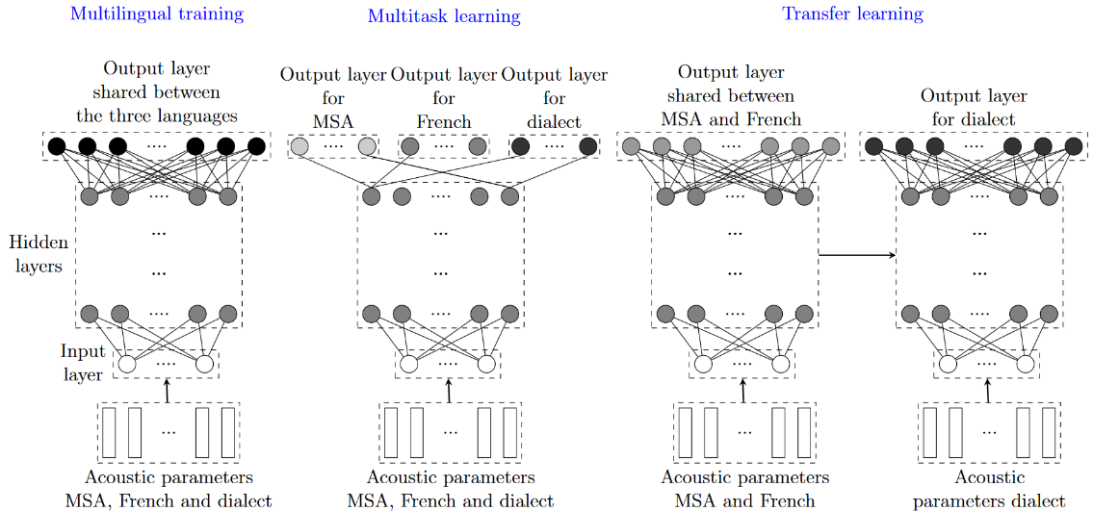


Figure 1.Proposed training techniques for the dialect acoustic modelling.

5.2. Multilingual training

In this approach we merged all the acoustic data of the three languages to construct a larger corpus. We then used it to train a TDNN without any distinction between the three languages. In this case, all the layers of the neural network are shared between the languages. Two questions were raised before the implementation of this solution:

- Knowing that MSA and French languages share some phonemes (e.g. /k/, /z/, etc.), how to find the best phonetic representation since the output layer that predicts triphones is shared between the three languages?
- How to optimize the necessary amount of speech data of each language (MSA and French) to make the contribution of each of them more effective on the performance of the ASR system of the Algerian dialect.

Concerning the first question, the integration of MSA and French data was carried out according to the two following approaches:

- **Union of phonemes** we simply take the union of the French and the MSA phonemes lists. This led to a set of 65 phonemes (34 MSA and 31 French).
- **Shared phonemes** the shared phonemes set is produced by keeping only one instance for each common phoneme. This led to a set of 47 phonemes (17 phonemes are common between MSA and French).

Concerning the optimization of the amount of data of each language, we decided to increase the training part of ADIC gradually byfew hours of each language(MSA and French) until reaching a total of 44 hours and then we select the combination that performs better on the Dev part of ADIC.

Figure 2 indicates the evolution of the Word Error Rate(WER) while adding at each step 4 hours of French data. The number above each curve represents the amount of MSA data (in hours). The blue and the black plots represent respectively the evolution of the WER when using the union of phonemes and when using shared phonemes. The WER in the baseline system (without adding MSA nor French data) is 30.05%. The best results (a WER of 28%) is the one got by adding 12 hours of MSA data and 12 hours of French data (see the curve (d)).

The experiments show that when increasing considerably the amount of MSA and French data (more than 12 hours), the results decrease. This last remark was expected, but we learned from these experiments the exact amount of the data necessary for improving the WER.

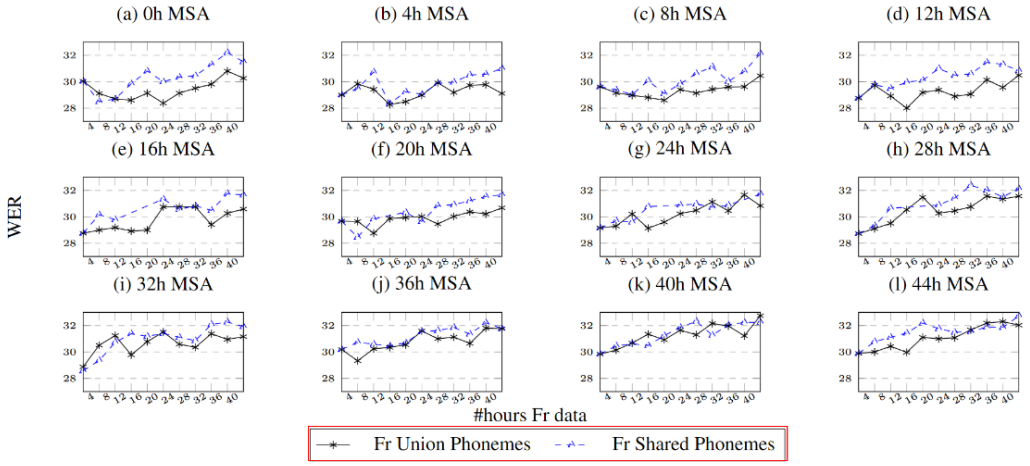


Figure 2. The WER variation on the ADIC Dev corpus for gradually extending ADIC Train corpus by MSA/French acoustic data.

5.3. Multitask learning

The principle idea of the multitask learning is to train one neural network with several sources of data to handle several tasks. For our purpose, we used the data of the three languages to train one model that recognises the three languages. Unlike the previous approach where all the layers of the neural network are shared among the three languages, in the multitask learning each language has a specific output layer, which means that the phonemes of each language are modelled separately. To update the parameters of the neural network, a simple way is to train it over different mini batches from each language. However, since we were not interested in the recognition of MSA nor French, the parameters of the neural network were updated by associating a weight w_l for each language in such a way that $\Sigma_{\{l=1\}}^{3} w_l = 1$ as applied in [21]. These weights were used to adjust the parameters of the hidden layers (λ_{sh}) after training over 400k samples of acoustic parameters according to the formula1.

$$\lambda_{sh} = \sum_{l=1}^{3} w_l \lambda_{sh}^{l} \qquad 1$$

This is equivalent to train three models separately one for each language where each model has a set of parameters corresponding to the shared layers λ_{sh}^{l} and those of the output layer λ_{out}^{l}. The parameters of the shared layers in the global modelλ_{sh} correspond to the weighted shared parameters of each model for each language.

To estimate the weightsw_l for each language and in order to give more importance to the dialect ASR task, our training started with a high dialect weight (0.8) and it decreased gradually with a step of 0.1 in a such way that the dialect has always the high weight comparing to MSA and French. We opted for this approach because the training process is time consuming and it is hard to explore all the searching space. At the end of the estimation process, the weights that ensure better results on the Dev ADIC were found to be 0.5 for dialect and 0.5 for MSA if the model is trained on two tasks. If the system is trained on three tasks, the value of those weights wasfounded to be 0.4 for dialect, 0.3 for MSA and 0.3 for French.

5.4. Transfer learning

In the case where a small amount of data is available to train the neural network, it is common to pretrain a model on a large dataset and use it as a fixed feature extractor for the new task. In this case, hidden layers of the original network were fixed and a new task-specific layers were added over them. As in the multitask learning, phonemes between languages are not shared in this approach, but we need to find a way to update the model parameters. The common used approach is to update the parameters of the new added layers using a large learning rate (0.0005 in our case) and to fine-tune the parameters of original hidden layers with a small learning rate (0.00005 in our case).

To estimate the number of hidden layers n to transfer, an initial neural network was trained on MSA and French data by using our multilingual training approach. Afterwards, the output layer of this network is replaced by a specific layer for the dialect while keeping the n-first hidden layers. Those are $n \in \{1,2,3,4,5\}$ knowing that the initial model is composed of 6 hidden layers. The obtained WER on the Dev part of ADIC are presented inFigure 3.The results show that keeping the four first hidden layers from the initial model ensures better results.

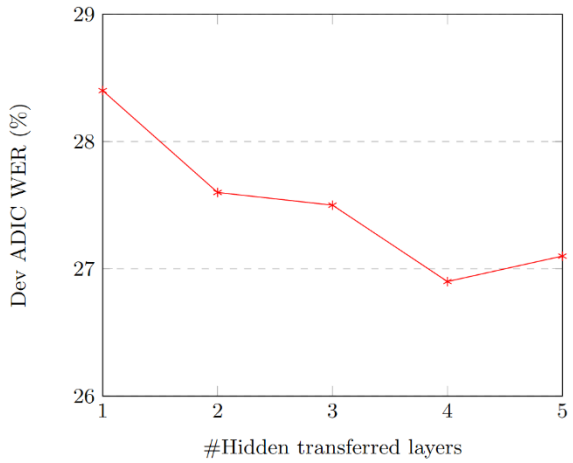


Figure 3. The impact of the number of hidden layers on the transfer learning.

6. Results and Discussion

We used several corpora to train the acoustic model: MSA spoken data were extracted from NEMLAR[2] and NetDC[3] corpora, French data were extracted from ESTER corpus [22], and the dialect data are our recorded corpus ADIC. Forty dimensional MFCC feature vectors are used as input of the neural network at each timestamp. These MFCC features are extended with 100-dimensional identity vector (i-vector) [23]. I-vectors are low-dimensional vectors of speech segment used to describe the speaker characteristic in the speech. Despite that technique was initially proposed for speaker verification and speaker recognition tasks, it is also useful for speech recognition since it encapsulates the speaker relevant information in a low-dimensional representation. The implementation was based on Kaldi [24], a state-of-the-art toolkit for speech recognition based on weighted finite state transducers[25], and the experiments were carried out on Grid5000 platform [26].

[2] http://catalog.elra.info/product_info.php?products_id=874
[3] http://catalog.elra.info/product_info.php?products_id=13&language=fr

6.1. Recognising the Algerian dialect using MSA-ASR system

We aim through this investigation to show how the Algerian dialect differs from the MSA. No dialectal data are used to train the language nor the acoustic models. The acoustic model is trained on 44 hours of MSA spoken data. We interpolate two bigram LMs trained on the MSA version of Gigaword corpus and on the transcripts of MSA speech training data; the interpolation weights are estimated on a MSA development corpus. The lexicon contains the most frequent words of the textual data used to train the LM. It has 95k unique words and 485k pronunciation variants. The results obtained with this ASR system are reported in Table 5. The test on MSA has been achieved on 5 hours of MSA speech data, while 1 hour and 15 minutes of the Test part of ADIC have been dedicated to test the ASR system on dialectal data.

Table 5. Performance of the MSA-ASR system on MSA and on the Algerian dialect.

System	Test	WER (%)	OOV (%)
MSA-ASR	MSA	12.7	2.5
	Test ADIC	76.3	33.6

Whereas the MSA-ASR system performs well on MSA (a Word Error Rate (WER) of 12.7%), it collapses completely when it is tested on the dialectal corpus (a WER of 76.3%). The Out-Of-Vocabulary (OOV) rate shows how MSA and Algerian dialect are different. These results confirm that it is impossible to directly recognise the Algerian dialect with an ASR system developed for MSA.

We report in the Table 6 the recognition results when applied on the Test part of ADIC according to the proposed approaches.

Table 6. The Algerian dialect speech recognition results according to the way of using data from foreign languages.

Training Approaches	Training data			WER (%)	OOV (%)
	Acoustic	Lexicon	Textual		
Monolingual training	44hMSA	MSA	MSA	76.3	33.6
	4hDial	Dial	Dial	42.6	7.9
	4hDial	Dial+MSA	Dial+MSA	39.7	6.8
Multilingual training	4hDial+44hMSA	Dial+MSA	Dial+MSA	36.6	6.8
	Union 4hDial+44hMSA+44hFr			36.3	
	Shared 4hDial+44hMSA+44hFr			37.1	
	Union 4hDial+12hMSA+12hFr			35.9	
	Shared 4hDial+12hMSA+12hFr			38.5	
Multitask learning	4hDial+44hMSA (mini batch)	Dial+MSA	Dial+MSA	36.5	6.8
	4hDial+44hMSA (weights averaging)			37.0	
	4hDial+44hMSA+44hFr (mini batch)			36.6	
	4hDial+44hMSA+44hFr (weights averaging)			36.9	
Transfer learning	44hMSA (Initial model) => 4hDial	Dial+MSA	Dial+MSA	38.1	6.8
	44hMSA+44hFr (Initial model) => 4hDial			37.1	

6.2. Monolingual training

In this approach, we used data from only one language to train the dialectal acoustic model. We remark that the use of dialectal data to train the different model improves the WER of the MSA-ASR system by 33 points (76.3% vs. 42.6%). These results were expected because of two main reasons: firstly, we used data that was specific to our task and thus led to a low OOV rate. Secondly, the dialectal phonemes were well-modelled by the acoustic model. Better still, including MSA textual data in the language modelling improves the system by 2.9% (42.6% vs. 39.7%).

6.3. Multilingual training

The multilingual training approach aimed to take advantage from the speech data of other languages to improve the recognition of the Algerian dialect. The experiments started by integrating the MSA spoken data in the training process of the acoustic model. This leads to an absolute improvement of 3.1% (39.7% vs. 36.6%) which shows how the MSA data are important in the acoustic and the language modelling of the dialect. However, integrating French data in the training process of the acoustic model of the Algerian dialect does not improve the WER. Even this poor improvement, we can remark that better results were obtained when the common phonemes between languages are modelled separately without any sharing. This could be explained by the fact that the shared phonemes between the MSA and the French languages, even if they are the same, they are used in different phonological contexts that makes their pronunciations different in each language. Consequently, they should be separated in order to ensure a good ASR system performance for the Algerian dialect. We also find that optimizing the size of the MSA and French acoustic data leads to a better result (a WER of 35.9%). This allow us to prevent the overfitting issue on MSA and French data.

6.4. Multitask learning

The neural network in the multitask learning was trained on several speech recognition tasks while allowing for the hidden layers to be shared and each task to have a specific output layer. We investigated two configurations according to the number of tasks to train. The first configuration aims to study the impact of MSA data on the speech recognition of the Algerian dialect by training the model on two tasks: ASR for MSA and for dialect. In the second configuration, the French ASR task was integrated in the training process.

For each configuration, we found that training the neural network over different mini batches from each language gives better resultscompared to the approach where we attributed weights for the different languages (weights averaging in Table 6). Knowing that the success of the weights averaging approach depends on the good estimation of the weights $w_l$for each language, we can explain the obtained results by our algorithm used to estimate these weights.In fact, we fixed a heigh weight for the dialect compared to the other languages;it would be interesting in this case to explore a large searching space where we fixed a low weight for the dialect.

The results also showthat training the neural network on two tasks (ASR for MSA and for dialect) leads to an absolute improvement of 2.7% compared to the case where the neural network was trained on one task (ASR for dialect). This shows the importance of MSA data on the acoustic modelling of the Algerian dialect and confirms the obtained results in the multilingual training. However, we found that integrating the French ASR task in the training process brings no benefit to the system's performance.

6.5. Transfer learning

We aimed in the transfer learning to train initial models on MSA and/or French data, to retain the four first hidden layers and to add new dialect task specific layer over those.

We trained two neural networks using the multilingual training approach to study the impact of each language on the ASR of Algerian dialect. The first model was trained only on the MSA data while in the second one the French data were also integrated in the training process. These two models are used, afterwards, for the transfer learning.

Unlike what we found in the multilingual training and in the multitask learning, the French spoken data improve the system's performance. The best WER was the one obtained by adapting the acoustic model trained on MSA and French data to the Algerian dialect.

By comparing our three approaches of integrating data from several languages into the training process of the acoustic model of the Algerian dialect, we found that the best approach is the one based on the multilingual training (a WER of 35.9%) where all layers of the neural network were shared between the three languages. This allows an implicit increase in the size of the data we used to train our model, which allows the model to better capture the relationship between the three languages and thus improve the dialect ASR system. It should be noted that the confidence interval for the system trained on the dialectal acoustic data only was ±1.65% (the one achieved a WER of 39.7% in Table 6), which means that integrating MSA and French spoken data in the training process of the acoustic model for the dialect achieves a significant improvement.

There is relatively few research works on ASR for Algerian dialects in order to be able to compare our obtained results. However, in the last edition of the MGB challenge, MGB5[27], there was a task about ASR for Moroccan dialect, which is relatively close to the Algerian dialect because they share several linguistic and acoustic aspects. The best system obtained a WER of 37.6%, knowing that 13 hours of dialectal speech were used with 1200 hours of MSA to train the acoustic model. This shows how hard the speech recognition task of Maghrebi dialects, especially the Algerian one, and that the results of our system are acceptable.

7. CONCLUSIONS

This work investigated developing an ASR system for Algerian dialect by starting from an ASR system dedicated to MSA. This attempt collapses completely when it was used to recognize the dialect (a WER of 76.3%). This shows how different are Algerian dialects and MSA.

To overcome the lack of spoken resources for this dialect and since Algerian dialects are mainly impacted by MSA and French languages, we investigated the use of acoustic data from these two languages. We showed that it is possible to develop an acoustic model on the base of a small recording dialectal corpus then adding it to larger corpora of well-resourced languages such as French and Arabic. It could be interesting to investigate this approach to develop ASR systems for other dialects especially those impacted by French such as Moroccan and Tunisian dialects. The recorded dialectal corpus provides a valuable resource for further studies on the Algerian dialect.

Through our investigation, we showed that sharing all layers of the neural network based acoustic model (the multilingual training) ensures best results compared to sharing only hidden layers (multitask and transfer learning). Furthermore, taking the union of phonemes of the three languages, in the case where the output layer is shared, led to a better acoustic model compared to the case of considering the intersection of common phonemes. We also investigated the required

amount of data required to train a decent dialectal acoustic model. Our conclusion is that selecting subsets of data led to a better speech recognition system compared to using a larger amount of data. This is because with larger amounts of speech data from one of the mixture of languages, the performance can be impacted negatively. The over representation of a particular language makes the ASR system more sensitive to the phonemes of this language and less to the others.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of Chist-Era for funding part of this work through the AMIS (Access Multilingual Information opinionS) project.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (seehttps://www.grid5000.fr).

REFERENCES

[1] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," Speech communication, vol. 56, p. 119–131, 2014.

[2] F. de Wet, N. Kleynhans, D. Van Compernolle and R. Sahraeian, "Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems," South African Journal of Science, vol. 113, p. 1–9, 2017.

[3] V.-B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: application to Vietnamese language," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, p. 1471–1482, 2009.

[4] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, L. Lamel, H. Maynard, M. Müller and others, "A very low resource language speech corpus for computational language documentation experiments," arXiv preprint arXiv:1710.03501, 2017.

[5] D. Amazouz, M. Adda-Decker and L. Lamel, "Addressing Code-Switching in French/Algerian Arabic Speech," in Proceedings of Interspeech, 2017.

[6] k. Abidi, M. a. Menacer and K. Smaili, "CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube," in 18th Annual Conference of the International Communication Association (Interspeech), 2017.

[7] M. Killer, S. Stuker and T. Schultz, "Grapheme based speech recognition," in Eighth European Conference on Speech Communication and Technology, 2003.

[8] H. Cucu, L. Besacier, C. Burileanu and A. Buzo, "Investigating the role of machine translated text in ASR domain adaptation: Unsupervised and semi-supervised methods," in 2011 IEEE Workshop on Automatic Speech Recognition Understanding, 2011.

[9] P. Karanasou and L. Lamel, "Comparing SMT methods for automatic generation of pronunciation variants," in International Conference on Natural Language Processing, 2010.

[10] S. Harrat, K. Meftouh, M. Abbas and K. Smaïli, "Grapheme to phoneme conversion - an Arabic dialect case," in Spoken Language Technologies for Under-resourced Languages, 2014.

[11] A. Masmoudi, F. Bougares, M. Ellouze, Y. Estève and L. Belguith, "Automatic speech recognition system for Tunisian dialect," Language Resources and Evaluation, vol. 52, p. 249–267, 01 3 2018.

[12] D. Yu, B. Varadarajan, L. Deng and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," Computer Speech & Language, vol. 24, p. 433–444, 2010.

[13] K. Kirchhoff and D. Vergyri, "Cross-dialectal acoustic data sharing for Arabic speech recognition," in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.

[14] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas and K. Smaili, "Machine translation experiments on PADIC: A parallel Arabic dialect corpus," in Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 2015.

[15] K. Meftouh, S. Harrat and K. Smaïli, “PADIC: extension and new experiments,” in 7th International Conference on Advanced Technologies ICAT, Antalya, 2018.
[16] K. Abidi and K. Smaïli, “An automatic learning of an Algerian dialect lexicon by using multilingual word embeddings,” in 11th edition of the Language Resources and Evaluation Conference, LREC 2018, 2018.
[17] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel and J. Glass, “A complete KALDI recipe for building Arabic speech recognition systems,” in Spoken Language Technology Workshop (SLT), 2014 IEEE, 2014.
[18] V. Peddinti, D. Povey and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[19] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, “Phoneme recognition using time-delay neural networks,” IEEE transactions on acoustics, speech, and signal processing, vol. 37, p. 328–339, 1989.
[20] K. Veselỳ, A. Ghoshal, L. Burget and D. Povey, “Sequence-discriminative training of deep neural networks.” 2013.
[21] R. Sahraeian and D. V. Compernolle, “Using Weighted Model Averaging in Distributed Multilingual DNNs to Improve Low Resource ASR,” Procedia Computer Science, vol. 81, pp. 152-158, 2016.
[22] S. Galliano, G. Gravier and L. Chaubard, “The ester 2 evaluation campaign for the rich transcription of french radio broadcasts,” in Proceddings of Interspeech, Brighton (United Kingdom), 2009.
[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, p. 788–798, 2010.
[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, “The KALDI Speech Recognition Toolkit,” in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.
[25] M. Mohri, F. Pereira and M. Riley, “Speech recognition with weighted finite-state transducers,” in Springer Handbook of Speech Processing, Springer, 2008, p. 559–584.
[26] D. Balouek, A. CarpenAmarie, G. Charrier, F. Desprez, E. Jeannot, E. Jeanvoine, A. Lèbre, D. Margery, N. Niclausse, L. Nussbaum, O. Richard, C. Pérez, F. Quesnel, C. Rohr and L. Sarzyniec, “Adding Virtualization Capabilities to the Grid'5000 Testbed,” in Cloud Computing and Services Science, vol. 367, I. I. Ivanov, M. van Sinderen, F. Leymann and T. Shan, Eds., Springer International Publishing, 2013, pp. 3-20.
[27] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals and K. Choukri, “The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech,” 2019.

INTEGRATED SPECIFICATION OF QUALITY REQUIREMENTS IN SOFTWARE PRODUCT LINE ARTIFACTS

Mworia Daniel, Nderu Lawrence and Kimwele Michael

Department of computing, Jommo Kenyatta University of Agriculture and Technology, Kenya

ABSTRACT

There are many calls from software engineering scholars to incorporate non-functional requirements as first-class citizens in the software development process. In Software Product Line Engineering emphasis is on explicit definition of functional requirements using feature models while non-functional requirements are considered implicit. In this paper we present an integrated requirements specification template for common quality attributes alongside functional requirements at software product line variation points. This approach implemented at analytical description phase increases the visibility of quality requirements obliging developers to consider them in subsequent phases. The approach achieves weaving of quality requirements into associated functional requirements through higher level feature abstraction method. This work therefore promotes achievement of system quality by elevating non-functional requirement specification. The approach is illustrated with an exemplar mobile phone family data storage requirements case study.

KEYWORDS

Software Product Line Engineering, Functional and Non-functional requirements, Quality attributes, feature variability, integration and requirements specification.

1. INTRODUCTION

In the history of requirements engineering, non-functional requirements (NFRs) were not considered alongside functional requirements until recently. NFRs problems are grouped into definition problems, classification problems, and representation problems. Representation of NFRs is a big challenge owing to their fuzzy nature where depending on how we define an NFR; its representation on a software specification document can make it appear like a functional requirement creating even more confusion in requirements documentation.

Despite the fact that there are many on-going efforts to determine in which stage of software development to integrate NFRs, researchers agree that taking NFRs into consideration during the early phases of any software engineering processes can improve the quality and agility of software[1].

There are various ways in which NFRs can be represented depending on the reason of their use and phase of the software development project. Goal-oriented approaches have advanced well-defined approaches to model NFRs at early stage of the requirement engineering process while at the architectural phase NFRs associated with particular components can be used to justify alternative designs [2].

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP, SOFE - 2021
pp. 91-106, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110309

Another very well-defined approach for representing NFRs is textual representation which involves documenting requirements in software requirement specification (SRS) through the use of templates. The most widely used textual requirements representation methods are the natural language-based templates [3].

In any software development process non-functional requirements (NFRs) analysis will yield performance requirements, business constraints, and non-functional properties or quality attributes (QAs). This work will focus on quality attributes requirements representation in software product line engineering (SPLE) where variability is critical and the operationalization of quality goals is closely interlaced with functional requirements.

In Software product line Engineering (SPLE) requirements engineering activities are carried out in the early stages of domain analysis & engineering (DA&E). A product-line is a set of products that share a common set of requirements, but also exhibit significant variability in requirements. In the requirements analysis stage, the requirements gathered in the previous stages are analysed and further refined. The commonalties and variabilities can be identified either by using product line specific techniques or other techniques such as feature-oriented domain analysis (FODA) and family-oriented abstraction, specification, and translation (FAST) [4].

SPLE exploits the similarities of the systems that belong to a product line and systematically handles the differences between them. Product line variability defines how product line applications may differ in terms of features, functional and quality requirements they fulfil. Like commonalities, product line variability is pre-planned by defining whether a given feature, functional or quality requirement is product line variability or not based on explicit decisions from all product management stakeholders [5] .Quality attribute variability can be due to functional variability causing indirect variation in qualities, and vice versa.

Most SPLE approaches typically cover the domain and application engineering processes, but set aside one activity important to companies which is analysis of non-functional properties (NFPs) or quality attributes and the evolution of SPL's artifacts. The large part of most SPL methodologies is management of functional variability and the minor part of implementing quality variability is with annotations that are sometimes abandoned after a short period of time because of the lack of integration during the SPL development activities.

Literature review clearly demonstrates the aspect of variability in quality attributes has been "neglected or ignored by most of the researchers and attention mainly put in the functionality variability of the products. As observed in[3], most approaches to quality attributes incorporation in software product line development introduce the variability at the design level (e.g., within sequences diagrams) instead of modeling the variability of the Quality attributes earlier on in the development process, such as the requirement level or at the architectural level. Our approach addresses this gap by considering and integrating quality attributes at the domain requirements analysis and specification phase.

Feature Models are the most widely used variability language, that model variability by means of high level features that are close to requirements specification. During feature model analysis it is important to consider quality attributes as part of the model variability alongside functional features to generate more than one solutions , the variation points be made explicit and document the decision models with the knowledge necessary to ponder about the better solution for each product to be derived. This work therefore proposes an approach that will support identification and integration of quality attributes with the functional features at respective variation point levels during domain requirements analysis phase based on higher-level abstraction of common features among variants.

The contribution of this paper is to provide support, applying domain analysis and variability management techniques, to the identification and representation of quality requirements in SPL development. This paper focus on analysis and specification of quality requirements alongside the functional requirements in the early stages of SPL development taking as input the domain requirement documents together with feature diagrams. The approach proposes the use of a textual integrated requirements template to extract common functional and quality attribute requirements at the SPL variation point. Further the Proposed approach extends the feature based analysis of domain requirements by focusing on the product family variation points to generate common functional quality attributes among the product family variants which are then stored as aspectual components to promote reuse.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed approach while Section 4 applies the conceptual approach on a case study. Finally, Section 5 summarizes the contributions of this work and outlines directions for further research.

2. RELATED WORK

A range of research works have been carried out that seeks to support incorporation of NFRs in software product line Engineering (SPLE) process. Whereas there is no agreed upon stage of integrating NFRs in to the software development process efforts in literature focus more on the solution space (design , architectural choice, evaluation and testing) than the problem space(requirements elicitation and analysis) [6].Some of the relevant approaches addressing this issue are presented below.

2.1. Quality attributes Integration based on Extension of UML models

In an effort to capture variability of quality attributes using Model-Driven Development (MDD), [7] recommend annotating the base model by means of extensions to the base modeling language. They add generic annotations related to a quality attributes like performance to the UML model which represents the set of core reusable domain assets. The concrete UML annotations are based on UML profiles with stereotypes to achieve desired quality attributes modeling. However annotations of the application base model prevents its reuse as well as that of derived quality attributes.

FeatuRSEB is a popular approach which combines FODA and the Reuse-Driven Software Engineering Business (RSEB) method . In FeatuRSEB UML-like notational constructs are used for creating Feature Diagrams, with explicit representation of variation points, and variants and explicit graphical representation for feature constraints and dependencies. Non-functional requirements are captured as feature constraints. Product Line Use Case modeling for System and Software engineering (PLUSS) is an approach that borrows from FeatuRSEB to combine Feature Diagrams and Use Cases. This approach makes explicit decomposition of the operator to compose a feature by introducing two new types of nodes; single adapters (represent XORdecomposition) and multiple adapters (OR decomposition).This approach however does not explicitly handle quality attributes.

2.2. Architectural based Quality Attributes integration Approaches

Extension of the feature model mechanisms from ATAM (Architecture Trade-off Analysis Method) can be used to represent quality attributes, their variability with respect to optionality and levels, their influence on quality of the functional, architectural and implementation features (indirect variation). The extended feature model presents both functional and quality concerns as

the fundamental element used to capture the variability in subsequent phases of design and implementation.

At the architectural phase existing works address quality attributes variability jointly with the variability of base applications. [16] propose the RiPLE-DE (RiSE Product Line Engineering - Design Engineering) process approach presenting the variability of quality attributes in feature diagrams and in order to derive the desired quality attributes the diagrams are enhanced with information of the base application (e.g., the system's response measure). The variation of attributes is presented in form of numerical values that will be used in evaluation of the resulting architecture designing SPL architectures that involve systematic transformation of functional requirements and quality.

Quality-driven Architecture Design and quality Analysis (QADA) is a method for incorporating attributes into software architectures, which do not however explicitly consider quality. Another approach in [8] suggest the influence of each feature on a non-functional property be predicted before generating the configurations. Their approach however focuses on predicting the effects of the features on individual applications instead of focusing on recurrent quality attributes at the domain engineering phase to promote reuse.

2.3. Goal Oriented NFR integration Approaches

[9] conclude that there is an association between software product lines and goal analysis and thus one can use goal-driven requirements approaches for feature specification. Goal analysis modeling can support auto-generation of feature models in SPLE. In SPLE paradigm, an integrated modeling framework (F-SIG, Feature-Softgoal Interdependency Graph) extends the feature modeling with concepts of goal-oriented analysis. This goal oriented analysis is aimed at letting developers to capture design rationale of inter-dependencies between variant features and quality attributes during the design of product line architecture, and evaluate the impact of variant features selected for a target system.

The goal driven and Chung's NFR framework approach has been widely used by researchers to integrate NFRs into the software development process. Whereas functional requirements are considered as hard goals, non-functional requirements are presented as soft goals in the analysis specification process. The correlation is shown as a directed graph where the nodes are hard goals, the target nodes are soft goals and the edges are represented by the + or – characters However software developers pay more attention to functional needs of a software and NFRs such as performance, usability, reliability and security are usually handled later in an ad-hoc manner mainly during the system testing phase [10].

NFRs can be essential in all aspects of Software Product Line (SPL) like in situations where a requirement may cuts across all product lines and the variation exists in the contextual application. [10] recommended extending the Product Line Use Case modeling for System and Software engineering (PLUSS) to include other NFRs other than the performance NFRs only by use of discrete values to express degree of satisfice-ability and for security NFR represent the levels of data protection as outlined in the NIST standard. This approach however focuses on how single NFRs can be evaluated for satisfiability during product testing.

[11] also proposes an approach of modeling quality attributes with the variability of the base application based on domain experts' judgments using the Analytic Hierarchical Process (AHP). This captured quality knowledge of domain experts is used for quality aware product. Any functionality that affects quality attributes is referred to as a contributor but do not explicitly deal with the quality attributes.

[12] advanced another interesting approach known as Concern-Oriented Reuse (CORE) ,a general- purpose software development which leverages on the strength of Model-Driven Engineering (MDE), Component-Based Software Engineering (CBSE), SPL, feature oriented and aspect-oriented software development, and goal modeling to promote reuse . This approach entails encapsulating all software functional and non- functional Requirements in reusable units called concerns. As much as they do not explicitly deal with quality attributes, the encapsulation of concerns is what our proposed approach recommends. The other difference with our work is the fact that they model variability of the component interfaces and not integration of functional and quality attribute concerns like our proposal suggest.

2.4. Domain Requirements Analysis and Specification

Our paper focuses on the textual representation of quality attributes alongside functional requirements in software product line so as to support documentation and subsequent phases of development. We therefore mention related works in the line of textual analysis and representation of quality attributes alongside FRs both in SPLE and single-system development approaches.

According to [13] the most common approaches for analysis and specification for software product lines can be categorized as product based specification, where the features of each individual product are specified one by one and feature based specification, where individual features are specified without links to any other features. There is also the family based specification approach where specification can be written for all the features of the product line with variable parts for individual features. Our approach to SPLE specification is similar to the family based specification with variable parts for individual features presented in a text based specification method.

Whereas [14] note that software product lines do not have a de facto standard for requirements analysis and specification there have been several attempts that promote to connect goal-oriented approaches with this task. [9] observe that feature modeling is the core of software product line engineering and a de facto standard in modeling variability in SPL.

Extended feature models can address representation of domain Quality attributes (such as performance, availability, security or safety) including their variation dimensions. This work extends this approach by considering the quality attributes variability alongside the functional variability at variation points. Existing requirements documentation methods separate functional and quality attribute requirements whereas at the variation point there could be common variation to all possible family members that could be integrated and documented together as aspectual components for easier reuse.

Volere Requirements Specification Template is a well-established method for recording requirements in a structured way. The method supports the recording of user goals and requirements in the template according to their rationale, associated stakeholder, priority and contextual details. There are different templates for specific NFRs like usability, maintainability security among others in the Volere documentation.

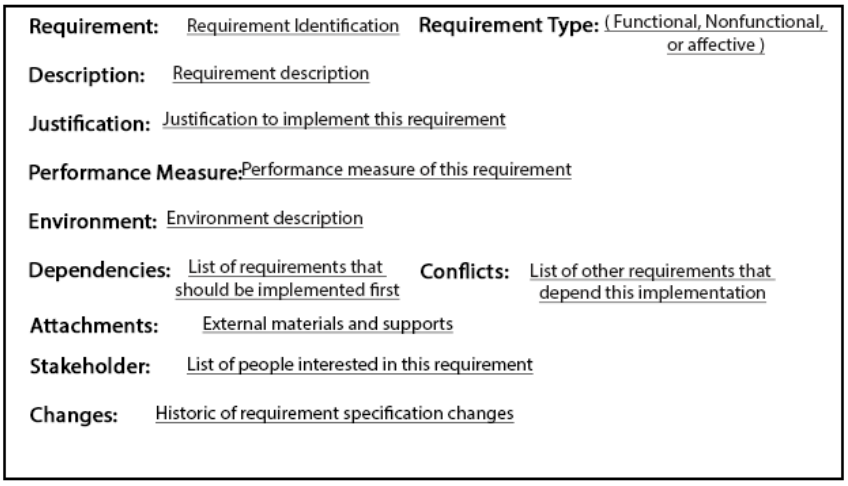


Figure 1. Volere Requirements Specification Template.

The Volere Requirements Specification Template documentation inspired several other works including [15] and [16]. A problem with the usage of such templates is that they are useful only when a single person is responsible for managing them. However, in a project where many people are working simultaneously, this can lead to inconsistent, contradicting and omitted requirements, and a need for a complex requirements management tool.

Apart from detailed tabular templates and models, several research works provide boilerplates (reusable sentences); a term referring to limited vocabulary sentences having specific placeholders to be completed in order to obtain semi-formal requirement sentences. [17]have presented an elicitation methodology by the use of their Non-functional Requirements Templates (NoRTs), which focuses on using generic statements(having core and optional parts) that become defined NFRs after adding required information. EARS approach provides a simple boilerplate for requirement templates that can be used for non-functional requirements as well.

18] use natural language processing techniques for identification of NFRs from requirements documents. The approach uses a language model and popular keywords for identification of NFRs. This work suffers from the limitation of the lexicon or keywords as most NFRs are domain dependent.

There have been different proposals for templates to support textual use case descriptions of Software Product Lines where fine-grained variation could be specified at the end of the SPL use cases with a template consisting of the following elements; name, type, line of the use case (the target of the variation), and description.

Another textual use case template found in [19] aimed at specifying the variation points through OPT and ALT tags where any text fragment of the textual use case description may be variant is explicitly marked by pairs of the XML-like tags <variant> and </variant>. [20] proposed a simpler tag notation where the tags are used only for marking variation points in use case scenarios of SPL. Each tag is expanded in a section called “Variations” and is mapped to the Orthogonal Variability Model (OVM).

[5] further observe organizations can also use their own specification templates or some standardized Software Requirements Specification (SRS) document structures to specify product line requirements. In order to capture the integrated quality attribute requirements at variation

points we propose to use own specification templates for documenting each variation point based on structured document templates such as extensible markup language (XML) which allow hierarchical representation of common and variable requirements.

It is clear from the literature review that existing SPLE specification models mainly focus on feature models, use cases and domain specific requirements specification languages. These approaches represent functional and non-functional requirements in separate documents and diagrams but our proposed approach recommends integrated specification document based on structured document templates such as extensible markup language (XML). The aspectual component development of the extracted functional quality requirement concern and the XML documentation can be handled using existing techniques in SPLE researchThe research works included in this section can be summarized as follows:

a) Existing models to integrating quality attributes into SPL development process do it more in the solution space (design , architectural choice, evaluation and testing) than the problem space(requirements elicitation and analysis).
b) All SPLE approaches discussed in related work above support analysis of quality attributes in respect to evaluation of achievement degree of non-functional property(NFP) in the final product but do not address the variation analysis of the quality attributes at the product family variation points.
c) Most of the text-based tabular templates represent quality attributes as independent elements of requirements process. The need for NFRs' relationship with specific functional requirements is not fulfilled by most of these efforts.
d) This work therefore focuses on textual extraction and integrated representation of functional quality attributes at respective variation points during domain requirements analysis phase based on higher-level abstraction of common features among variants. The Functional quality attributes can the then be included in requirements documents to achieve traceability and incorporation throughout the development process.

3. PROPOSED APPROACH

With ever increasing number of software development companies adopting Software Product Line (SPL) methodology as opposed to single-systems software development the field is also continuously changing. In terms of the process many approaches typically cover the domain engineering activities of variability modeling but ignore issues that matter to organizations such as the analysis of non-functional properties (NFPs) or quality attributes and the evolution of SPL's artifacts[7]). A few organizations that attempt to implement NFP variability do so with annotations that are sometimes abandoned after a short period of time because of the lack of integration among the SPL activities.

A persistent challenge in SPL development has been the modeling and management of variations in their product lines. In SPL development variability exists at different levels of abstraction, including requirements variability (mainly feature based), architecture variability (mainly component based), and implementation variability (mainly code based). In most modern software systems variability can also be classified as variability in functional behaviour, variability in non-functional system properties and fault based variability. This work focuses on requirements variability and possible integrated specification of functional and quality requirements in the early phases of software product lines development.

From domain knowledge & stakeholders requirements documents, in the feature oriented analysis phase we can extract common functional quality attributes among the variants and use higher level feature abstraction method to map them to respective variation points as common

base concerns of an SPL. The Steps in the integrated specification of requirements at the SPL variation points are as follows.

3.1. Identification of variability point of interest from an SPL feature model

An established way of capturing commonalities and variability's of a product-line system during early development stages is by use of Feature-Oriented Domain Analysis (FODA). In FODA a feature model presents requirements in a tree based feature diagram where functional features are decomposed into more fine-grained features that are either mandatory, optional, or alternative and optional features specify variability.

Since variability analysis determines where variability is needed in the product line and features represent system property or functions relevant to some stakeholder, a product family variation point could also yield members which also present common and variable requirements limited by the domain scope. This work therefore pursues the possibility of creating a requirement specification template that can support the integration of common functional and quality attributes in a software product line variation point. The integrated requirements can be stored as an aspectual component for reuse.

Product family variability is where a feature can have alternative implementations or variant implementations, which can be chosen to create different products. A variation point is each point in the software where different variant implementations from a variant population can be chosen from. Characteristics of a product that can be changed to produce a different product are called variation points. In terms of realization technique, a variation point can be the point where a class is chosen to be used or where code fragments are chosen to be run. Once you identify the related variant features of the product family in the graph a variation point can be marked with every set of related features [21].

3.2. Analysis of requirements at the variation point

One way to incorporate the non-functional requirements early in development of SPLs is to consider them at the variation point where common and variable features among the different variants can be analyzed to identify common functional and non-functional properties. The requirement commonality and variability applies to both functional and non-functional aspects with respect to family members. However there exists common quality factors that are associated with functional requirement in each domain such as security for banking systems, reliability for embedded systems and usability in general for most of the applications . Early identification of such requirements can facilitate development of reuse of aspectual components to serve all members of a product line family.

At each variation point the core and possible functional and non-functional requirements of the family members can be identified and analysed according to a structured specification template. Whereas non-functional requirements can be classified as Performance, quality and constraints requirements, our focus is quality properties that would assure end user satisfaction. NFRs and especially quality attributes have a close relationship with functional requirements especially in their operationalization. We therefore propose creating a relationship of certain quality factors with FRs as fulfilling the quality factors at the variation points level eventually supporting the NFRs satisfaction at the global level.

A major objective of Software product line engineering is maximizing the commonalities (platform or architecture) whilst minimizing the cost of variations (i.e., of individual products) to facilitating reuse in a predictive manner. Whereas several methods and tools were developed for

variability identification exist such as FODA , that are specifically focused on requirements, including Feature-oriented domain analysis (FODA) , Natural Language (NL) requirements documents can also act as a source of variability information that can used to define variability models [22].

Since a high level requirements can hide a family of different products this work pursues variability extraction based on analysis of high level requirement documents or features which when further decomposed can yield common quality attribute operationalization among the variants. These integrated requirements are functional quality attributes.

3.3. Integrated specification of functional and quality requirements to produce an aspectual component for the SPL

Existing SPL specification models mainly focus on feature models, use cases and domain specific requirements specification languages. These approaches represent functional and non- functional requirements in separate documents and diagrams. Whereas functional decomposition is done in feature diagrams and use cases the associated non-functional requirements are not explicitly defined. This work proposes a semi-formal approach using structured non-mathematical notations to organize information about functional quality requirements.

As [3] observed there are situations where a non-functional requirement affects neither a single functional requirement nor the system as a whole but a specific set of functional requirements. Such a case requires unique variability specification templates that ensure explicit documentation and adequate explicit traceability. The specification template can then be developed using the aspect oriented design methodology for reuse among the family members.

At the Variation point of a feature model we identify a dominant functional concern and decompose the system model hierarchically into sub- features that contribute to its realization. We also identify a core non-functional property (Quality attribute) of the domain at the variation point and refine it into specific quality concerns for each possible family member.

As argued in [23] even a cross cutting non-functional requirement within a software family such as security may differ in intensity levels such as intense security or moderate level of security. Non- functional requirements can be grouped into performance requirements, constraints and quality attributes. Whereas Constraints such as cost, efficiency or portability can be mapped to architectural or implementation decisions, quality attributes such as security , usability and error handling can be mapped directly to functional components and thus referred to as functional quality attributes (FQAs). These FQAs are normally required by several applications in a product line and therefore specialized components can assure their satisfaction.

An integrated textual requirements analysis template as in Table 1 can generate possible functional and non-functional requirements at the variation point exposing common functional quality attributes which apply to all members of the software product family with a common base at that variation point. The common functional-quality attribute set of requirements can then be stored as an aspectual component at the variation point of the primary feature model through a join relationship and the same can be applied at every level of variation point that presents similar characteristics.

Table 1. Elements of proposed variation point integrated requirements template.

Variation point(Vp) Id	(description)
VpFunctional Requirements	(description)
VpQuality Requirements	(description)
VpQualityConcern	(description)
VpFunctional-QualityConcern	(description)

The integrated functional-quality requirement becomes the final requirement description that will be used in all subsequent phases of software development including design decisions .This inclusion of quality attributes in functional description will remind the developers to consider them in all decisions and subsequent phases of software development [24].

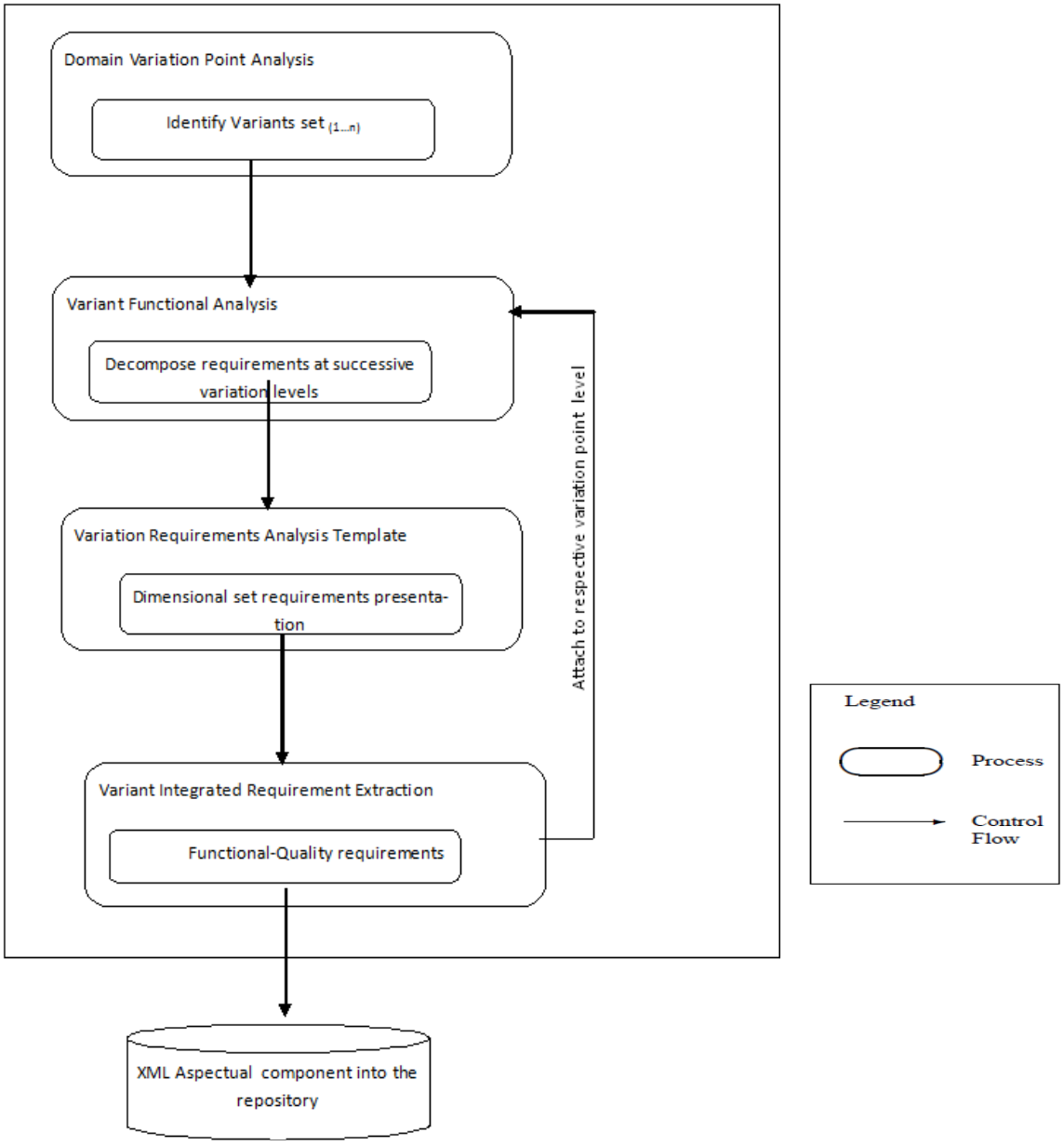


Figure 2. Proposed approach for integrating quality concerns at SPL variation point.

4. CASE STUDY

For practical demonstration of the proposed approach we present case study that consists of a simplified version of variability requirements from a mobile phone software product line family. Customizable software is necessary for a broad spectrum of domains (e.g., operating systems for diverse hardware) and hence our choice of mobile phone family data storage features programming.

Modern mobile phones are multifunctional and provide the ability to perform a wide range of actions beyond the common voice communication role. Common mobile phone features and utility functions include log in, call management, text messaging, storage, camera ringtones clock, and varying multimedia features. Among increasingly critical functions of a mobile phone is data storage which cannot only be extended with flash memory card device but also with online backup. Phones as storage devices hold personal, organizational and even proprietary data.

Research findings consistently show that a significant portion of mobile phone users are concerned about security of their mobile device, its data, or its application against a "casual" and unprofessional attack by children, spouses, friends, co-workers etc. Implementing this security feature for different members of the mobile phone family requires variability management in terms of functionality and quality attributes of the system. We will focus on variability of the phone data protection and user privacy enforcement mechanisms as requirements that expose functional quality attributes at the variation points.

i) Identifying variation point dimensions

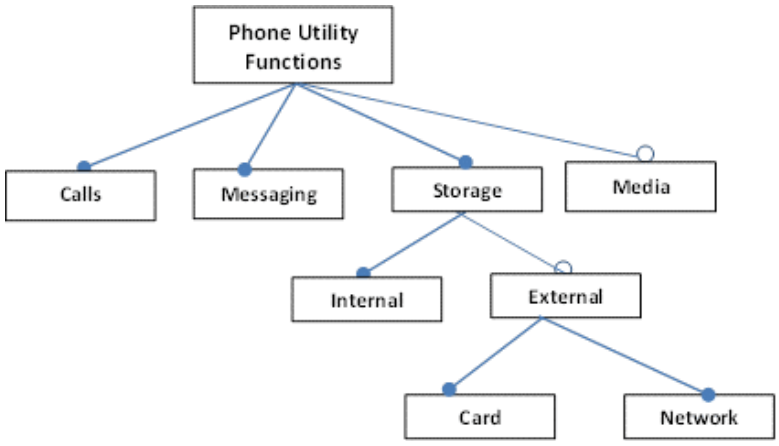


Figure 3: Mobile phone utility functions feature diagram

ii) Requirements analysis and specification at variation point

From Figure 3 we focus on storage feature as a critical function in mobile phones today since they are being used as personal digital assistants (PDAs) for private and even corporate work. Variants in the mobile phone family line will present different abilities to satisfy that storage function. Assuming the following set of general user expectations from the phone family line expectations related to data storage:

Rq1.The phone shall have capacity to store data
Rq2. The phone shall have ability to extend storage capacity
Rq3. The phone shall have capacity to clear storage once full
Rq4. The phone may (optionally) permit transfer data to other devices
Rq5. The phone shall have capacity to read different file formats
Rq6. The phone shall ensure security of data
Rq7. The phone shall ensure user privacy

Rq6 and Rq7 are non-functional requirements and specifically quality requirements which must be achieved by all variants to some level of satisfaction through different mechanisms. Addressing the satisfaction of the two quality requirements involve consideration of functional quality attributes at respective family tree variation points.

At the level of domain analysis the requirements above will subject our feature graph to further functional decomposition to identify different phone capabilities to operationalize them with different mechanisms and limitations. The broad techniques of achieving the requirements is at the phone login, desktop, database and external interface points as shown in Figure 4 with further variability among the possible solutions.

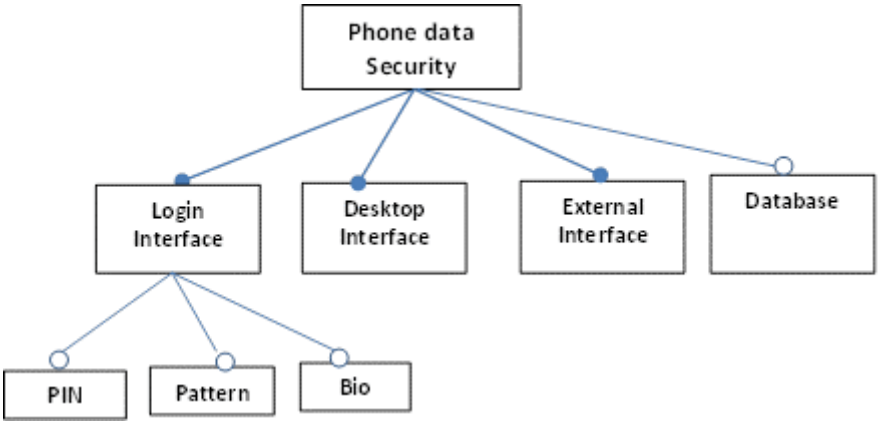


Figure 4. Mobile phone data security requirements feature diagram

iii) Integrated specification of Functional quality attribute requirements

Assuming we have three variants of the phone family that differ in their ability to satisfy the requirements Table 2 illustrates the possible scenarios:

Table 2: Functional quality achievement analysis matrix

VariantType	**Ability to Satisfy**
Smart	Rq1,Rq2,Rq3,Rq4,Rq5
Evolving	Rq1,Rq2,Rq3,Rq4
Dumb	Rq1,Rq3

In order to support integrated specification of common functional quality requirements at the variation points we need to analyse the variants further with respect to ability and quality attribute satisfaction mechanisms.

Looking at the security feature implementation capabilities for the different variants at different data access interface points a domain features function analysis template can generate the common functional quality requirements as shown below:

Table 3: Functional quality achievement analysis matrix

Phone Variant	**Login Interface**			**Desktop Interface**			**External Interface**		
	Pass	Bio	Patt	Pass	Bio	Patt	RW	H/w key	Enc
SMART	X	X	X	X		X	X	X	X
EVOLVE	X				X	X	X		X
DUMB	X				X				

Nb. Symbol (X) in the matrix denotes the variant supports the associated security achievement mechanism, Pass(Password), Bio(Biometric) RW(Remote wipe), H/w(Hardware , Enc (Encryption) and Patt (pattern).

Table 3 presents an analysis template that indicates satisfaction of security and privacy quality requirements in the three variants phone data storage function happens in three dimensions (at login, Desktop and External interfaces). However the mechanisms of satisficing the quality requirements generate both common and variable mechanisms possible in the product line as follows:

At Login security variation point, all the three variants share the PIN authentication mechanism of access control, while some support pattern, biometrics or both.

For Desktop security/privacy point all the three variants share auto- screen lock access control but differ in unlocking mechanism of PIN, pattern, biometrics and key- combination.

For External Storage security variation point two variants share remote wipe and encryption capabilities but one has hardware key and the other does not have the functionality.

The analysis above therefore generates three functional-quality requirements at the variation points as shown in Table 4.

Table 4. Functional quality achievement analysis matrix.

Variation Point	**Functional- Quality requirement**	**Specification ID**
Login Interface	Authenticate-PIN	VPlogin-Auth(PIN)
Desktop Interface	Display Lock –Auto/key lock	VPDesk-Lock(KEY)
External Interface	Encrypt	VPExt-Auth(encrpt)
External Interface	Remote wipe	VPExt-protect(Rw)

iv) Storage in the repository inform of XML aspectual component

The four common functional-qualities attributes (FQAs) for the three variants at different variation points can thus be developed separately as aspectual components to be attached to the common base architecture at respective join points defined by variation points. In order to make the requirements specification systematic and traceable the functional quality attributes can be stored in XML format in the repository together with the original SRS documents for future reuse.

This approach supports the architects and application engineers while generating new members or variants of the software product line family that is initially restricted by defined scope.

5. DISCUSSION AND CONCLUSIONS

In this paper, we suggested a practical approach for integrating functional quality requirements in SPL requirements documentation in an intuitive way. We have outlined steps in the process of analysis and integration and demonstrated practicality of the proposed approach with a case study.

The proposed approach is based on domain feature model analysis and natural language textual representation, which is the most widely, used methods in SPLE. Literature review shows a lot of variability analysis in functional dimensions while quality variability is considered implicit. Our approach therefore supports early consideration of quality attributes and their subsequent integration into the SPL documentation.

Since natural language and textual description of software requirements can be used to extract functional features and identification of variation points is a continuous activity in all phases including requirement gathering, this work attempted to extract quality attributes variations during analysis that can be represented alongside functional requirements owing to their means of operationalization. This work however is limited to incorporation and representation of quality attributes whose realization is based on functional view of software.

One limitation in this work is the fact that it has not been tested in a complete product line architecture that specifies the rules on how the aspectual components will be connected as well as their relationships, interactions, and dependencies among them. For example very elaborate security quality component implementation can negatively affect usability attribute and cost objectives. We therefore hope to investigate these scenarios in an industrial scope.

State of the art solutions to modern day problems demands automation which has not been accomplished in this work. To encourage adoptability of this approach we intend to develop a tool to manage automated extraction of variation point functional quality attributes using natural language processing techniques and latent semantic analysis abstraction from software product line family requirements documentation.

In future, we aim to explore the impact of weaving quality attributes to functional requirements at variation points considering that quality attributes have conflicts and interdependencies with others.

REFERENCES

[1] R. R. Maiti (2016). Capturing, Eliciting, and Prioritizing (CEP) Non-Functional Requirements Metadata during the Early Stages of Agile Software Development. Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, College of Engineering and Computing. (968) https://nsuworks.nova.edu/gscis_etd/968.

[2] L. Chung, B. A. Nixon, E. Yu,. & J. Mylopoulos (2012). Non-functional requirements in softwareengineering (Vol. 5). Springer Science & Business Media.

[3] G. Carvalho, F. Barros.and A. Sampaio A (2015). "NAT2TEST tool: From natural language requirements to test cases based on CSP." Software Engineering and Formal Methods. Springer, Cham, 2015. 283-290.

[4] J. M. Horcas ., M. Pinto & L. Fuentes (2019). Software Product Line Engineering: A Practical Experience. In 23rd International Systems and Software Product Line Conference - Volume A (SPLC '19), September 9–13, 2019, Paris, France. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3336294.3336304

[5] A. Metzger, & K. Pohl, (2014). Software product line engineering and variability management: Achievements and challenges. FOSE. 10.1145/2593882.2593888

[6] J. M. Horcas (2018). WeaFQAs: A Software Product Line Approach for Customizing and Weaving efficient Functional Quality Attributes. A Doctoral Dissertation at the university of University of Malaga, Spain. Accessed online from http://orcid.org/0000-0002-7771-0575

[7] R. Tawhid, & D. C. Petriu, (2011) Automatic derivation of a product performance model from a software product line model," in 15th International Software Product Line Conference, ser. SPLC, 2011, pp. 80{89. 21

[8] N. Siegmund, M. Rosenm, Muller, C. Kastner, Giarrusso, P. G., S. Apel, and S. S. Kolesnikov, (2013). Scalable prediction of non-functional properties in software product lines: Footprint and memory consumption," Information & Software Technology, vol. 55, no. 3, pp. 491{507, 2013. [Online]. Available: https://doi.org/10.1016/j.infsof.2012.07.020 24, 26

[9] F. Q. Khan, S. Musa, & G. Tsaramirsis (2018). A novel requirements analysis approach in SPL based on collateral, KAOS and feature model. International Journal of Engineering & Technology, 7 (4.29) (2018) 104-108

[10] Nguyen, Q.L. (2009). Non-Functional Requirements analysis modeling for software product line. Proceedings of the 2009 ICSE Workshop on Modeling in Software Engineering, Washington, D.C., 56-61.

[11] G. Zhang, H. Ye, an& Y. Lin, (2014). Quality attribute modeling and quality aware product configuration in software product lines," Software Quality Journal, vol. 22, no. 3, pp. 365{401, Sep 2014. [Online]. Available: https://doi.org/10.1007/s11219-013-9197-z 20, 24, 26, 29, 78

[12] M. Schottle , O. Alam, J. Kienzle, & G. Mussbacher, (2016).On the modularization provided by concern-oriented reuse," in Companion Proceedings of the 15th International Conference on Modularity, ser. MODULARITY Companion 2016. New York, NY, USA: ACM, 2016, pp. 184{189. [Online]. Available: http://doi.acm.org/10.1145/2892664.2892697 20, 25, 29, 78

[13] F. Q. Khan, , S. Musa, & G. Tsaramirsis (2018). A novel requirements analysis approach in SPL based on collateral, KAOS and feature model. International Journal of Engineering & echnology, 7 (4.29) (2018) 104-108

[14] S. Chimalakonda & D. H.Lee. (2016). On the Evolution of Software and Systems Product Line Standards. SIGSOFT Softw. Eng. Notes 41, 3 (June 2016), 27-30. DOI: http://dx.doi.org/10.1145/2934240.2934248

[15] C.Porter, E. Letier, & M. A. Sasse, (2014, August). Building a National E-Service using Sentire experience report on the use of Sentire: A volere-based requirements framework driven by calibrated personas and simulated user feedback. In 2014 IEEE 22nd International Requirements Engineering Conference(RE) (pp. 374-383). IEEE.

[16] M. F. A.Carvalhaes, A. F. d.Rocha , A. M. F.Vieira & T. M. G. d.Barbosa (2014). Affective Embedded Systems: a Requirement Engineering Approach. International Journal of Emerging Trends & Technology in Computer Science 8(2):70-75. DOI: 10.14445/22312803/IJCTT-V8P113

[17] S. Kopczyńska & J. Nawrocki, (2014, August). Using non-functional requirements templates for elicitation: A case study. In 2014 IEEE 4th International Workshop on Requirements Patterns (RePa) (pp. 47-54). IEEE.

[18] M. Younas, K.Wakil, D. N. Jawawi, M. A., Shah & A. Mustafa, (2019). An Automated Approach for Identification of Non-Functional Requirements using Word2Vec Model. International Journal of Advanced Computer Science and Applications(IJACSA), 10(8), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100871

[19] I. S. Santos, R.M. Andrade, and P.A. Neto (2015).Templates for textual use cases of software product lines: results from a systematic mapping study and a controlled experiment.. Journal of Software Engineering Research and Development (2015) 3:5 DOI 10.1186/s40411-015-0020-3

[20] W. Choi, S. Kang, H Choi, , J. Baik (2008) Automated generation of product use case scenarios in product line development. In: Proceedings of the International Conference on Computer and Information Technology. IEEE Computer Society,Washington, DC, USA

[21] González-Huerta, J., Insfran, E., Abrahão, S. and McGregor, J. D., 2012. Non-functional requirements in model-driven software product line engineering. In Proceedings of the Fourth International Workshop on Nonfunctional System Properties in Domain Specific Modeling Languages - NFPinDSML '12. New York, New York, USA: ACM Press, pp. 1–6

[22] A. Fantechi,. , S. Gnesi, & L. Semini, (2019) From Generic Requirements to Variability. Accessed on 5/12/2020 from http://ceur-ws.org/Vol-2376/NLP4RE19_paper16.pdf

[23] J. Jean-Marc (2012). Model-Driven Engineering for Software Product Lines. Review Article in ISRN Software Engineering Volume 2012, Article ID 670803, 24 pages doi:10.5402/2012/670803

[24] M. A. Gondal, N. A. Qureshi, H. Mukhtar, and H. Ahmed,. (2020). An Engineering Approach to Integrate Non-Functional Requirements (NFR) to Achieve High Quality Software Process. In Proceedings of the 22nd International Conference on Enterprise Information Systems - Volume 2: ICEIS, ISBN 978-989-758-423-7, pages 377-384. DOI: 10.5220/0009568503770384

AUTHOR INDEX