IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2020, accepted January 26, 2020, date of publication January 29, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970178

A Novel Software Engineering Approach Toward Using Machine Learning for Improving the Efficiency of Health Systems

MOHAMMED MOREB[1], TAREQ ABED MOHAMMED[2], OGUZ BAYAT[1], AND OGUZ ATA[1]

[1]Graduate School of Science and Engineering, Altinbas University, 34217 Istanbul, Turkey
[2]Tareq Abed Mohammed, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk 36001, Iraq

Corresponding author: Mohammed Moreb (mahammed.moreb@ogr.altinbas.edu.tr)

ABSTRACT Recently, machine learning has become a hot research topic. Therefore, this study investigates the interaction between software engineering and machine learning within the context of health systems. We proposed a novel framework for health informatics: the framework and methodology of software engineering for machine learning in health informatics (SEMLHI). The SEMLHI framework includes four modules (software, machine learning, machine learning algorithms, and health informatics data) that organize the tasks in the framework using a SEMLHI methodology, thereby enabling researchers and developers to analyze health informatics software from an engineering perspective and providing developers with a new road map for designing health applications with system functions and software implementations. Our novel approach sheds light on its features and allows users to study and analyze the user requirements and determine both the function of objects related to the system and the machine learning algorithms that must be applied to the dataset. Our dataset used in this research consists of real data and was originally collected from a hospital run by the Palestine government covering the last three years. The SEMLHI methodology includes seven phases: designing, implementing, maintaining and defining workflows; structuring information; ensuring security and privacy; performance testing and evaluation; and releasing the software applications.

INDEX TERMS Health dataset analysis, machine learning, methodology, software development management, software engineering.

I. INTRODUCTION

The field of health informatics (HI) aims to provide a large-scale linkage among disparate ideas. Normally, a healthcare dataset is found to be incomplete and noisy; as a result, reading data from dataset linkage traditionally fails within the discipline of software engineering. Machine learning (ML) is a rapidly maturing branch of computer science since it can store data on a large scale. Many ML tools can be used to analyze data and yield knowledge that can improve the quality of work for both staff and doctors; however, for developers, there is currently no methodology that can be used. Regarding software engineering, there has been a lack of approaches to evaluating which software engineering tasks are better performed by automation and which require human involvement or human-in-the-loop approaches [1]. Big data has many challenges regarding analysis challenges for real-world big data [2], including OLAP mass data, mass data protection, mass data survey and mass data dissemination.

Recently, a set of frameworks have been used to develop data analysis tools such as Win-CASE [3] and SAM [4]. The market has vast data analysis tools that can discover interesting patterns and hidden relationships to support decision makers [5]. BKMR used the R package as a statistical approach on health effects to estimate the multivariable exposure-response function [6].

Augmentor included the Python image library for augmentation [7], while for the visualization of medical treatment plans and patient data, CareVis was used [8], as it was designed for this task. Other applications require a visual interface using COQUITO [9]. For health-care data analytics, the widely known 3P tools [10] were used. Many simple applications, such as WEKA, which provided a GUI for many machine learning algorithms [11], while Apache Spark was used for the cluster computing framework [12], are powerful

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Aljawarneh.

TABLE 1. Big data analytics tools according to the task.

Big Data Task	Tools
Data Languages	R, Python
Data Integration	Blockspring, Pentaho
Data Collection	Import.io
Data Cleaning	OpenRefine, DataCleaner
Data Analysis	Qubole, BigML, Statwing
Data Visualization	Tableau, Silk, Charito, Plot.ly
Data Mining	Rapid Miner, Weka, Oracle data mining, IBM Modeler
Data Storage and Management	Hadoop, MongoDB, Talend

systems that can used in various applications for solving problems using big data and machine learning [13]. Table 1 summarizes the main tools used for big data in analytics according with respect to the task. Software engineering for machine learning applications (SEMLA) discusses the challenges, new insights, and practical ideas regarding the engineering of ML and artificial engineering (AI) [14]. NSGA-II proposed algorithms for real-world applications that include more than one objective function for enhancing performance in terms of both diversity and convergence [15]. ML algorithms in clinical genomics generally come in three main forms: supervised, unsupervised and semi-supervised [16]. Interflow system requirement analysis (ISRA) has been used to determine the system requirements [17].

Electronic healthcare (eHealth) frameworks have replaced traditional medical frameworks to improve mobile healthcare (mHealth) and enable patient-to-physician and patient-to-patient interactions to achieve improved healthcare and quality of life (QoL) [18]. Big data and IoT have been used for improving the efficiency of m-health systems by predicting potential life-threatening conditions during the early stages [19]. Intelligent IoT eHealth solutions enable healthcare professionals to monitor health-related data continuously and provide real-time actionable insights used to support decision making [20].

Machine learning is a field of software engineering that frequently utilizes factual procedures to enable PCs to ''learn'' by using information from saved datasets. Unsupervised or information mining focuses more on exploratory information investigation and is known as learning supported by data analytics. Patient laboratory test queue management and wait time prediction are a challenging and complicated job. Because each patient might require different phase operations (tasks), such as a check-up, various tests, e.g., a sugar level test or blood test, X-rays or surgery, each task can consider different medical tests, from 0 to N, for each patient according to their condition.

In this article, based on a grounded theory methodology [21], the researchers proposed a novel methodology, SEMLHI, in developing a framework by defining the research problem and methodology for the developers. The SEMLHI framework includes a theoretical framework to support research and design activities that incorporate existing knowledge. The SEMLHI framework was composed of four components that help developers observe the health application flow from the main module to submodules to run and validate specific tasks. This enables multiple developers to work on different modules of the application simultaneously. The SEMLHI framework supports the methodological approach to conducting research on health informatics. It also supports a structure that presents a common set of ML terminology to use, compare, measure, and design software systems in the area of health. This creates a space whereby SE and ML experts can work on a specific methodological approach to enable health informatics software development teams to integrate the ML model lifecycle. Our methodology was applicable to current systems or in the development of new systems that use the ML module for current systems, which can be used in regular updates to add data to the system, to perform irregular updates and to add new features such as new versions of ICD diagnosis codes, minor model improvements for bug fixes, new functionalities required by the client, and new hardware or architectural constraints.

II. METHODS AND DISCUSSION

Based on original data collected from a hospital run by the Palestine government covering the past three years, first, the data were validated, and all outliers were removed. Then, the remaining data were analyzed using the developed framework to compare ML techniques that predict test laboratory results. Our proposed module was compared with three systems engineering methods: Vee, Agile and SEMLHI. The results were used to implement the prototype system, which requires a machine learning algorithm. After the development phase, a questionnaire was delivered to the developer to indicate the results of using the three methodologies. The SEMLHI framework was composed of four components: software, machine learning model, machine learning algorithms, and health informatics data. The machine learning algorithm component uses five algorithms to evaluate the accuracy of the machine learning models for various components.

We used the original data as the selected dataset to develop a patient prediction test laboratory result prediction model, and the patient was required to perform more than one test. In this article, we focus on helping patients and doctors complete their treatment tasks by using predictable test results based on the International Classification of Diseases [22] (ICD-10) and helping hospitals save time and reduce effort dedicated to medical testing. Using the SEMLHI framework, realistic patient data were analyzed carefully and rigorously based on important parameters such as age, start time, end time, patient treatment, and detailed treatment content for each task. We identified the laboratory tests required for patients based on their conditions and the operations performed during treatment. The patient data included only codified variables, including ICD-10 codes, procedure codes, and medication orders, often reduced to smaller subsets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2986259, IEEE Access

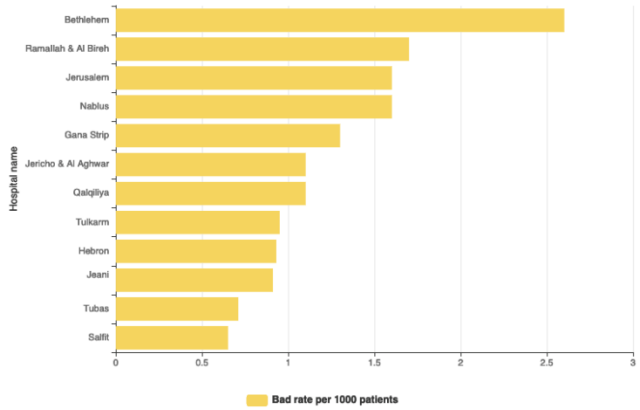

FIGURE 1. Bed rate for patients by city.

A. DATASET AND PREPROCESSING

The application delivery of applied ML models in healthcare was often hampered by the existence of isolated product deployments with poorly developed architectures and limited or nonexistent maintenance plans. The ''Translating Research into Agile Development'' (TRIAD) method presents a five-step method for designing a tailored EHR tool [23].

The SEMLHI models and methodology were developed by including new software systems connected to real datasets and presented knowledge from the data using ML algorithms to improve the efficiency of the required system. The dataset case studies discussed in this article were set within the context of Palestine hospitals and centers. Three hospitals and nine medical centers were used for our dataset. Figure 1 illustrates the summary of bed rats per 1k patients distributed across 12 cites. Furthermore, data collection was conducted over the last three years, and 458k patients were identified with corresponding patient nos. Overall, for the PMC dataset, 141k patients with 1.63% missing, a mean of 1.08M, a std dev of 554k, a min of 10k, and a max of 1.04M were included. For the age label, 141k patients with 1.63% missing, a mean of 32.24, a std dev of 26.25, a min of 0, a max of 88, and a median of 29 were considered.

B. AVAILABLE FEATURES OF PATIENTS

The patient dataset included 457914 cases and nine tables. Each table had different features, and many techniques could be implemented, such as semantic coordination for intelligent databases [24], feature selection problems using genetic algorithms [25], and new gene-weight mechanisms [26]. Some features were connected with other tables to build datasets describing the main attribute, as mentioned in the next section.

The laboratory test data include 200,000 cases (columns); each case has a basic attribute such as the patient no., gender, age, department, diagnosis code, description, and date of the lab test. Figure 2 summarizes the main features used on the sample dataset from all data, along with the distribution,

FIGURE 2. Main features used on the sample dataset.

TABLE 2. Comparison of the three system engineering methods, Vee, Agile and SEMLHI, proposed in this article.

	SEMLHI	Vee	Agile
Flexibility	very high	rigid	very high
Emphasis	risk	specification	customer
Logic	depth first	breadth first	depth first
Assumption	independent iteration	stable info	independent iteration
Scope	medium and large	large and complex	small
Iteration	very quick	slow	very quick
Delivery	one-shot delivery	one-shot delivery	incremental delivery

center value and dispersion. The data are represented as $L = \{l1, l2, l3, l4, \ldots, ln\}$, where l is the item of the laboratory test reports, and n is the number of items.

III. CONCEPTUAL FRAMEWORK AND DESIGN

A. SEMLHI METHODOLOGY

The SEMLHI methodology is used in software development in the health area. For traditional applications, the development process includes many methodologies, such as the waterfall methodology, spiral methodology, and agile methodology, which can be used to define and develop the software. Table 2 illustrates the results of the comparison between our methodology, the Vee [27] methodology, and the Agile [28] methodology.

The SEMLHI framework methodology describes in detail the process that was used when developing health software and the mechanism used to integrate and use ML algorithms with the development software. The SEMLHI methodology provides a developer with a new road map for designing health applications with system functions and software implementation. This framework includes ten stages, starting from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2986259, IEEE Access

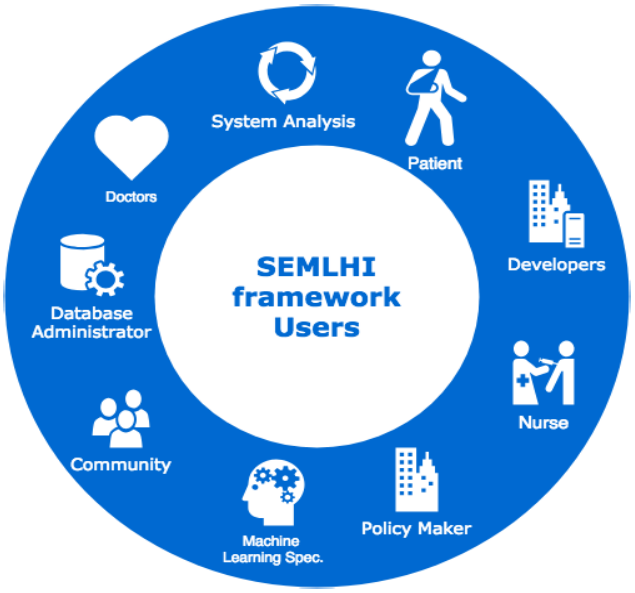

FIGURE 3. SEMLHI framework users.

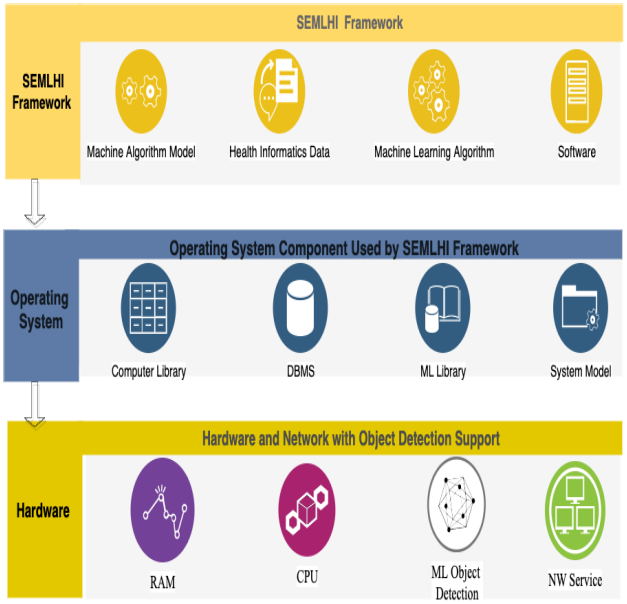

FIGURE 4. SEMLHI conceptual framework.

defining the problem until reaching the stage of development and ending with the results, as explained in the next section.

To develop the HI system, developers follow many sequence steps, such as design (encode data, define outliers and clean the data), implement (verification and validation), maintain defined workflows, structure information, provide security and privacy, test the performance, and then release the software applications. Records in most datasets in HI are weakly structured and non-standardized. To apply ML to the HI system, a set of patterns must be used by the algorithm to predict and visualize the ML algorithm and generate knowledge. The main patterns that were used in our framework were the geographic location, patient records, departments and hospitals, surgical history, obstetric history, family history, habits, immunization, assessment and plan, and test results. The next section describes the details of our framework, which is composed of four components.

B. SEMLHI FRAMEWORK

SEMLHI frameworks were specifically geared toward facilitating the development of software applications and include components that facilitate the analysis of a health dataset. Many users, as illustrated in Figure 3, will work directly as developers or system analysts with approach frameworks or indirectly by using the results. Figure 4 summarizes the proposed framework as a conceptual framework, in addition to the mechanism used to interact with the operating system and hardware.

For software engineers, our frameworks interact with operating system components that were used by the framework, and all software manages the device hardware with the main system device used by the framework.

Our framework was composed of four components or modules (software, machine learning model, machine learning algorithms, and health informatics data). Figure 5 shows how each module interacts with all modules to work as a framework.

FIGURE 5. SEMLHI framework components.

1) HEALTH INFORMATICS DATA

In ML, data are essential, and choosing the methods for presenting and visualizing knowledge is the most important step. Our dataset sample contains ten columns with 50k rows (cases). To use a dataset on health informatics

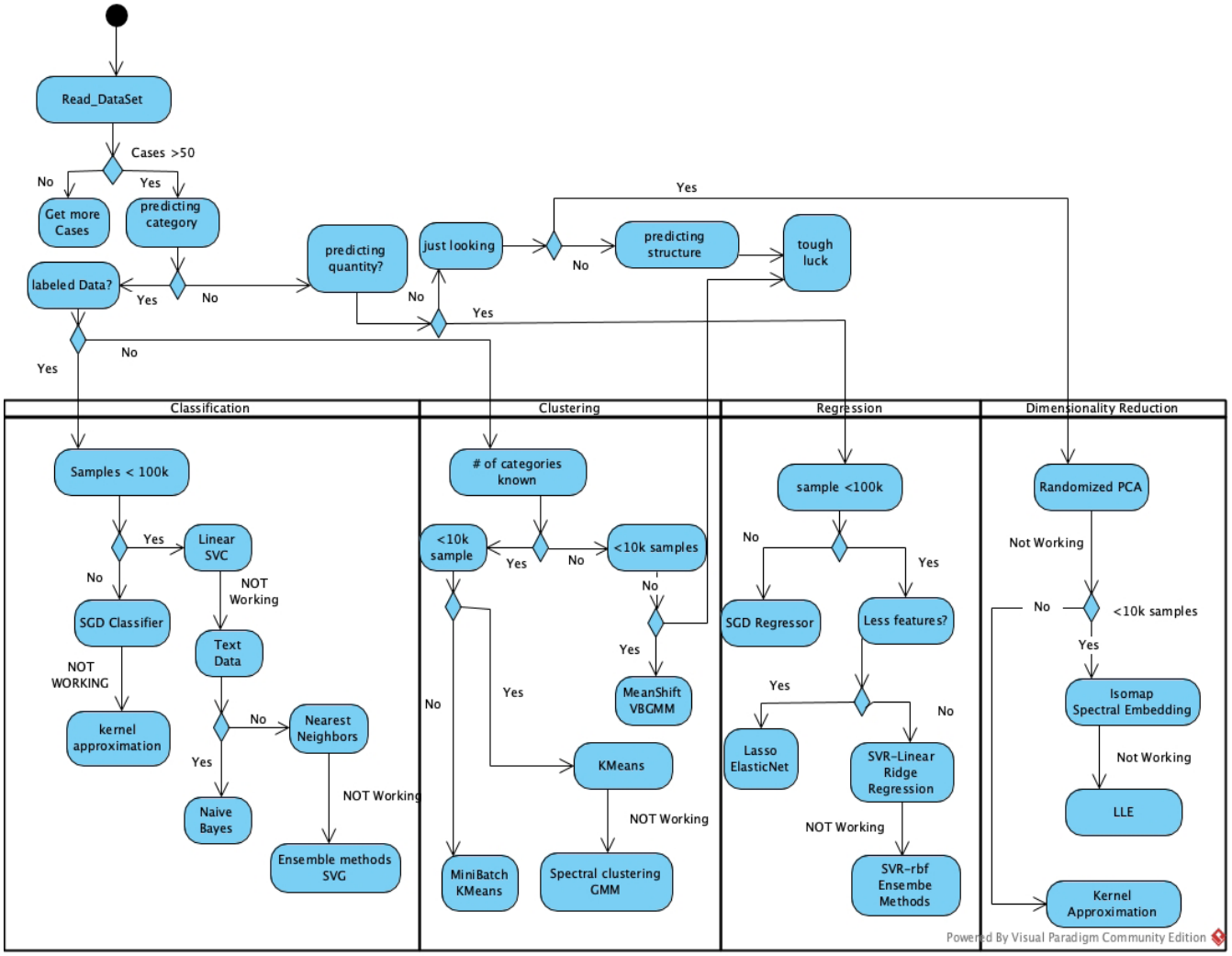

FIGURE 6. **Machine algorithm model components.**

data (HID) algorithms, a transformation into numerical features was required. Other data contain missing, duplicate or null values, such as negative ages and extremely large integers, which could negatively affect the performance of our ML algorithm. Figure 6 describes the main roles in detecting the methodologies used in the machine algorithm model, which are classification, clustering, regression and reduction.

HID uses data sources and a dictionary for translation during label encoding to convert each value in a column to a number to reduce the amount of misinterpreted data used by Bayesian inference. A node identifier was used to analyze data as a common process with patterns determined using patient-specific research identifiers. A dataset usually requires multiple records from the same patient to be identified as being related in the deidentified database. For outlier HID, a set of methods was used in the analysis to find hidden groups to remove outliers, and in an advanced step, the outlier values of the data that appear to be erroneous need to be found and cropped from the dataset. Addressing incomplete data in unsupervised clustering, chi-square and Fisher's exact tests were performed to determine the patterns that are discriminating between pair clusters [29].

To predict disease, we used ICD-10 with multiple labels, as each patient has an ICD code in their health records, which can affect all regions of the retina. However, there is currently no classification system [30] for distinguishing anterior (peripheral) and posterior (macular) data. We hypothesize that these classifications were characterized by D and refractive features, highlighting the disparity in the types of disease.

Collected electrocardiograph data were used to focus on the D most common diagnosis cases in the laboratory test result database: $D = \{d1, d2, d3, \ldots, d_n\}$, where d is a disease that was applicable to a diagnosis code and n is the number of disease classes using the k-means algorithm with multiple labels. Algorithm 1 presents the pseudocode for the k-nearest neighbor algorithm for multilevel learning.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2986259, IEEE Access

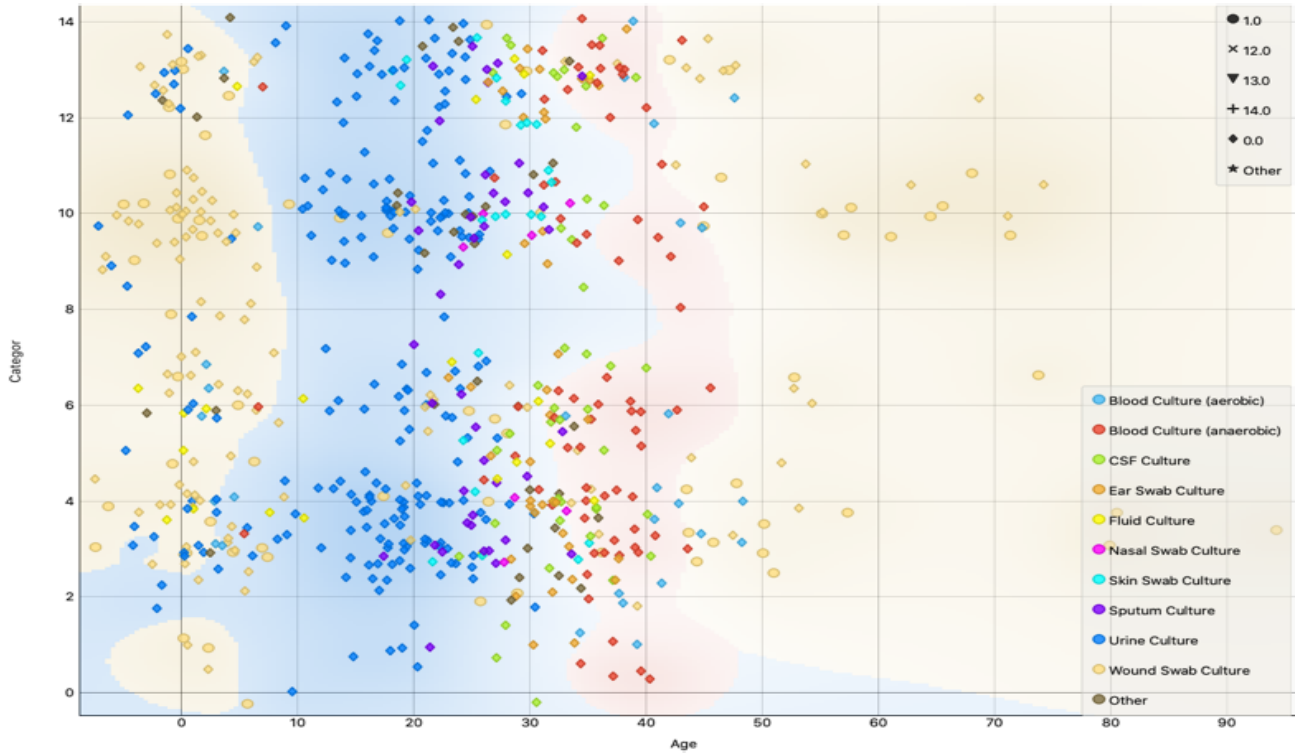

FIGURE 7. Age and category feature summary with other variables (test results, gender, and ICD-10).

Algorithm 1 k-Nearest Neighbor Algorithm for Multi-level Learning by Using Correlation, Diagnosis Code, and Label Weight With Frequency

input: Heterogeneous data source, number of k, correlation n. training set;
label of x
while *While condition* **do**
1- for all heterogeneous datasets, we will work on correlating multiple labels, adding one label for i = l to t and then joining the table for i = l to t, join table
2- apply ML to one label based on Freq. DG weight, and Lowes accuracy
3- create new role based on step 2
4- apply new role to all new schema; create new role if micro = sen, and if test = normal, then mml = 3
5- classify ML based on DC category
6. predict new disease based on role created;
end

The data module reads the data from data sources, such as CSV files or any other available sources; this module includes a set of algorithms that automatically remove missing values, clean the data to remove noise, and encode some features. Predicting missing values with incomplete data, for classification, normally requires decision trees; for a small amount of sample data or large numbers of genes, feature selection techniques, such as genetic algorithms and particle swarm optimization, are widely used [31]. The DRFLLS tool gave the best estimation to estimate missing values for a dataset that has a small rate of missing values [32].

As we have 750 case categories in our sample test data, represented by a 27-laboratory test, after running this module, a new dataset that includes 18 columns and 750 rows is generated. Figure 7 summarizes the ages with category features clustered by laboratory test results.

2) ML ALGORITHMS

Machine learning algorithms (MLAs) are used to compute the parameters that might define a model [14], optimize its network topology and improve the system convergence without losing information. MLAs including submodules are listed in Table 3.

As a supervised learning method, k-nearest neighbors (KNN) [33] can be used for classification and prediction problems. KNN makes decisions based on the dominant categories of k objects rather than a single object category. Figure 8 identifies most of the MLAs used for health classification.

As all the data in our sample of datasets were prepared using the SEMLHI framework, the output method will supervise the "label data" for this KNN algorithm with multiple labels and evaluate our result (KNN was used for supervised learning, while k-means was used for unsupervised learning). k-Means can be used for datasets that include a million labeled data points. Approximate nearest neighbors

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2986259, IEEE Access

TABLE 3. Machine learning algorithms sub model.

Sub Model for MLA Component	*Applications*	*Algorithms*
Classification	Spam detection, Image recognition	nearest neighbors, random forest, SVM
Regression	Drug response, Stock prices	SVR, ridge regression, Lasso
Clustering	Customer division, Grouping test outcomes	k-Means, spectral clustering, mean-shift
Dimensionality reduction	Visualization, Increased efficiency	PCA, feature selection, non-negative matrix factorization
Model selection	grid search, cross validation, measurements.	
Preprocessing	Transforming input information, for example, content for use with AI algorithms.	preprocessing, feature extraction.

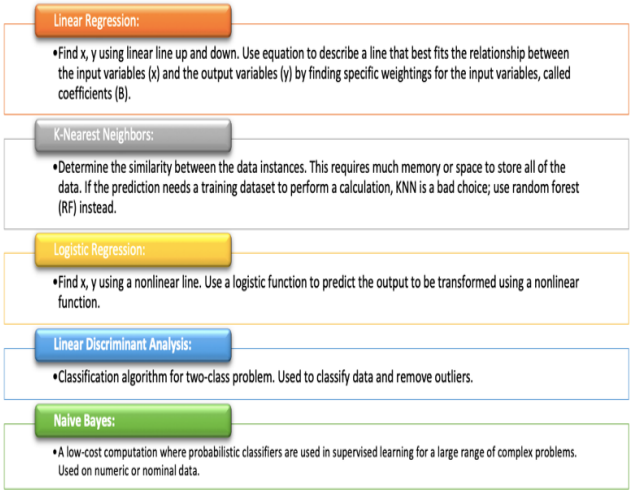

FIGURE 8. Machine learning algorithms used for health classification.

(ANNs), which is usually 10x - 100x faster than KNN support vector machines (SVMs), is a good and fast solution for many problems and will almost always outperform KNNs. Figure 9 shows that logistic regression has high accuracy compared with expected and real predictions.

In supervised learning, the dataset contains 'n' rows (cases); each case needs to be evaluated using a function $f : A \rightarrow B$ to compare with label A or label B according to the function 'f' by evaluating E and comparing them to learn from the training set of n. f has a set $n(d)$. In unsupervised learning, the data are not labeled. To apply the data in the

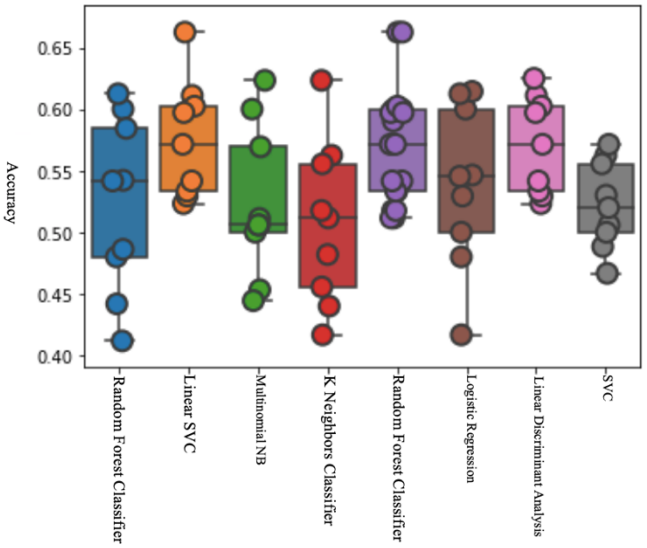

FIGURE 9. High accuracy of logistic regression compared with that of other algorithms.

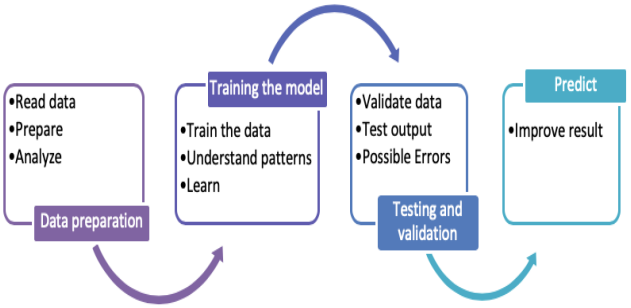

FIGURE 10. Mechanism of the machine algorithm model.

analysis, dimensional reduction is applied [34]. The training *set (t)* is used, which includes n objects $(t) = xi \in A: 1 \leq i \leq 0$ that can be in category n of class $C1$, $ck \in A$, by applying algorithm f in the evaluation phase to the set, where ck takes the input $x \in Cj: 1 \leq j \leq k$.

For clustering, we need to calculate the distance d between the two objects x and y by comparing the values of their n features and applying the Minkowski metric.

3) MACHINE ALGORITHM MODEL

Machine learning helps us extract useful features from a dataset to address or predict health-related events [35]. The machine algorithm model (MAM) component includes five submodules: read the data, prepare the data, train the model, test and evaluate the model, and predict new data. Figure 10 describes the sequence of these stages.

The challenge for this component was to use the right type of algorithm, which can optimally solve the dataset while avoiding high bias or variance. The main component of the MAM was used to analyze the dataset based on the set of conditions. If the dataset includes > 50 labeled samples, then classification algorithms will be used for the selection;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2986259, IEEE Access

TABLE 4. Evaluation of the accuracy of machine learning models.

Algorithm Name	*Accuracy*
KNN Classifier	0.487694
Linear SVC	0.564566
Logistic Regression	0.560412
Multinomial NB	0.517013
Random Forest Classifier	0.488955

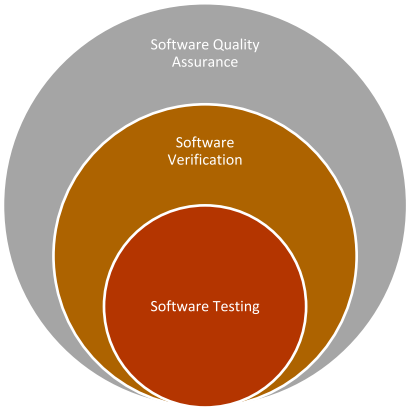

FIGURE 11. The software module includes subclasses including reuse, performance, testing, privacy, and security.

otherwise, cluster algorithms will be applied. If the dataset requires prediction of the quantity, regression algorithms will be used; otherwise, dimensional reduction will be applied.

Based on the original data, five algorithms were used to predict the laboratory test results utilizing the MAM component of the SEMLHI framework. ML approaches and algorithms can achieve better performances than expert-knowledge-based approaches [30]. ML algorithms use two types of techniques: supervised learning and unsupervised learning. For the MLA module, we first determine which techniques to use; then, we select the most suitable algorithms to use based on mathematical selection related to certain criteria. For the different algorithms applied, Table 4 shows their accuracy results (KNN classifier, linear SVC, logistic regression, multinomial NB, and random forest classifier).

We compared our approach with previously published systems in terms of performance to evaluate the accuracy of the machine learning models. The accuracy results for different algorithms were obtained after applying them to 750 cases, with linear SVC having values of approximately 0.57, compared with the KNN classifier, logistic regression, multinomial NB, and random forest classifier.

4) SOFTWARE

The software module, which is visualized in Figure 11, includes a subclass that includes reuse, performance, testing, privacy, and security. For software testing, the main point was to verify that the code was running correctly by testing the code under known conditions and checking that the results were as expected [36]. Visual analytic and interactive visualizations offer a higher degree of freedom for users for feature filtering, sorting patterns according to different interestingness measures, templating, and providing details on demand. Various visualization techniques, such as EventExplorer, ActiviTree, MatrixWave and DecisionFlow, can be used. Patterns can be clustered using an SOM or projection method, while plot patterns can use the double-decker method [37].

This class was used to test the memory or CPU resource usage for the application. The performance issues were determined by first measuring them and then profiling the code. Then, the optimization of that code was carried out using the benchmark, which was the best choice for comparing the results to improve the optimization performance. Code smells found genetic algorithms, used by 22.22%, to be the most commonly utilized machine learning techniques [38].

In multi-label classification, a prediction containing a subset of the actual classes should be considered better than a prediction that contains none of them, i.e., predicting two of the three labels correctly is better than predicting no labels at all. To measure a multi-class classifier, a misclassification using micro- and macro-averaging was carried out [16].

The security module has a significant impact on software development, maintenance, cost, and quality; security processes are implemented by integrating security activities and tools in the software development process, utilizing security requirement management, and providing training for developers.

IV. CONCLUSION

This article addressed an important HI with ML topic in software engineering by proposing an efficient new method approach related to software engineering, identified in prior research studies, using original data sets collected during the last 3 years from a Palestine hospital. This methodology allows developers to analyze and develop software for the HI model and create a space in which software engineering and ML experts can work together on the ML model life-cycle, especially in the health field. This manuscript proposed a framework that included a theoretical framework composed of four modules (software, ML model, ML algorithms, and HI data). The new methodology was compared between three system engineering methods: Vee, Agile and SEMLHI. The results showed the delivery of the new methodology for one-shot delivery. For the MAM component on the SEMLHI framework, laboratory test results were obtained using five algorithms to test the accuracy of the ICD-10 results using equations and to evaluate the accuracy of the ML models with a sample size of 750 patients. The results for MAM showed that the SVC was approximately 0.57.

AVAILABILITY OF DATA AND MATERIALS

Data that support the findings of this research were available from The Palestinian Ministry of Health, but restrictions were applied to the availability of these data, which were used under license for the current study and thus were not publicly available. Data are, however, available from the authors upon

reasonable request and with permission of the Correspondence Author.

ABBREVIATIONS

SEMLHI: Software Engineering for Machine Learning in Health Informatics
SE: Software Engineering
ML: Machine Learning
HID: Health Informatics Data
HI: Health Informatics
MAM: Machine Algorithm Model

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The research meets all applicable standards with regard to the ethics of experimentation and research integrity, and the following was certified/declared true. The informed consent of human participants was obtained in written format, and it was approved by The Palestinian Ministry of Health. As an expert scientist and along with coauthors in the concerned field, the paper has been submitted with full responsibility, following due ethical procedure, and there is no duplicate publication, fraud, plagiarism, or concerns about animal or human experimentation.

CONSENT FOR PUBLICATION

Not applicable.

COMPETING INTERESTS

All authors report no conflicts of interest.

FUNDING

Not applicable.

ACKNOWLEDGMENT

The authors would like to thank Health Minister of the State of Palestine, Dr. J. Awwad, for allowing us to access the Palestinian dataset for patients, and for all the teams that supported us during the last two years, the feedback from whom greatly improved this manuscript.

REFERENCES

[1] A. Holzinger, "Interactive machine learning: Experimental evidence for the human in the algorithmic loop," *Appl. Intell.*, vol. 49, no. 7, pp. 2401–2414, 2019.
[2] T. A. Mohammed, A. Ghareeb, H. Al-Bayaty, and S. Aljawarneh, "Big data challenges and achievements: Applications on smart cities and energy sector," in *Proc. 2nd Int. Conf. Data Sci., E-Learn. Inf. Syst.*, 2019, p. 26.
[3] B. Cakici, K. Hebing, M. Grünewald, P. Saretok, and A. Hulth, "CASE: A framework for computer supported outbreak detection," *BMC Med. Inform. Decis. Making*, vol. 10, no. 1, p. 14, 2010.
[4] A. J. Vickers, T. Salz, E. Basch, M. R. Cooperberg, P. R. Carroll, F. Tighe, and J. Eastham, and R. C. Rosen, "Electronic patient self-assessment and management (SAM): A novel framework for cancer survivorship," *BMC Med. Inform. Decis. Making*, vol. 10, no. 1, p. 34, 2010.
[5] A. Ismail, A. Shehab, and I. M. El-Henawy, "Healthcare analysis in smart big data analytics: Reviews, challenges and recommendations," in *Security in Smart Cities: Models, Applications, and Challenges*, vol. 9, A. E. Hassanien, M. Elhoseny, S. H. Ahmed, and A. K. Singh, Eds. Cham, Switzerland: Springer, Nov. 2019, pp. 27–45.
[6] J. F. Bobb, B. C. Henn, L. Valeri, and B. A. Coull, "Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression," *Environ. Health*, vol. 17, no. 1, p. 67, 2018.
[7] B. Aribisala and O. Olabanjo, "Medical image processor and repository–MIPAR," *Inform. Med. Unlocked*, vol. 12, pp. 75–80, Jul. 2018.
[8] W. Aigner and S. Miksch, "CareVis: Integrated visualization of computerized protocols and temporal patient data," *Artif. Intell.in Med.*, vol. 37, no. 3, pp. 203–218, Jul. 2006.
[9] J. Krause, A. Perer, and H. Stavropoulos, "Supporting iterative cohort construction with visual temporal queries," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 91–100, Jan. 2016.
[10] R. K. Pathinarupothi, P. Durga, and E. S. Rangan, "Data to diagnosis in global health: A 3P approach," *BMC Med. Inform. Decis. Making*, vol. 18, no. 1, pp. 1–13, 2018.
[11] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 438–441.
[12] Q.-C. To, J. Soto, and V. Markl, "A survey of state management in big data processing systems," *VLDB J.*, vol. 27, no. 6, pp. 847–872, Dec. 2018.
[13] S. R. Salkuti, "A survey of big data and machine learning," *Int. J. Elect. Comput. Eng.*, to be published. Accessed: Jan. 7, 2020. [Online]. Available: http://ijece.iaescore.com/index.php/IJECE/article/view/19184/pdf
[14] F. Khomh, B. Adams, J. Cheng, M. Fokaefs, and G. Antoniol, "Software engineering for machine-learning applications: The road ahead," *IEEE Softw.*, vol. 35, no. 5, pp. 81–84, Sep. 2018.
[15] T. A. Mohammed, Y. I. Hamodi, and N. T. Yousir, "Intelligent enhancement of organization work flow and work scheduling using machine learning approach tree algorithm," *Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 6, pp. 87–90, 2018.
[16] J. A. Diao, I. S. Kohane, and A. K. Manrai, "Biomedical informatics and machine learning for clinical genomics," *Hum. Mol. Genet.*, vol. 27, no. R1, pp. R29–R34, May 2018.
[17] P.-H. Cheng, Y.-P. Chen, and J.-S. Lai, "An interflow system requirement analysis in health informatics field," in *Proc. WRI World Congr. Comput. Sci. Inf. Eng.*, vol. 1, 2009, pp. 712–716.
[18] C. George, P. Duquenoy, and D. Whitehouse, "eHealth: Legal, ethical and governance challenges," in *eHealth: Legal, Ethical and Governance Challenges*, C. George, D. Whitehouse, and P. Duquenoy, Eds. Berlin, Germany: Springer, 2014, pp. 1–398.
[19] K. N. Mishra and C. Chakraborty, "A novel approach towards using big data and IoT for improving the efficiency of m-health systems," in *Advanced Computational Intelligence Techniques for Virtual Reality in Healthcare*, vol. 875. Cham, Switzerland: Springer, 2020, pp. 123–139.
[20] B. Farahani, M. Barzegari, F. Shams Aliee, and K. A. Shaik, "Towards collaborative intelligent IoT eHealth: From device to fog, and cloud," *Microprocessors Microsyst.*, vol. 72, Feb. 2020, Art. no. 102938.
[21] C. Oliver, "Critical realist grounded theory: A new approach for social work research," *Brit. J. Social Work*, vol. 42, no. 2, pp. 371–387, Mar. 2012.
[22] J. Disantostefano, "International classification of diseases 10th revision (ICD-10)," *J. Nurse Practitioners*, vol. 5, no. 1, pp. 56–57, Jan. 2009.
[23] K. D. Clark, T. T. Woodson, R. J. Holden, R. Gunn, and D. J. Cohen, "Translating research into agile development (TRIAD): Development of electronic health record tools for primary care settings," *Methods Inf. Med.*, vol. 58, no. 1, pp. 1–8, Jun. 2019.
[24] T. A. Mohammed, S. Alhayli, S. Albawi, and A. Deniz Duru, "Intelligent database interface techniques using semantic coordination," in *Proc. 1st Int. Sci. Conf. Eng. Sci.-3rd Sci. Conf. Eng. Sci. (ISCES)*, Jan. 2018, pp. 13–17.
[25] T. A. Mohammed, O. Bayat, O. N. Uçan, and S. Alhayali, "Hybrid Efficient Genetic Algorithm for Big Data Feature Selection Problems," *Found. Sci.*, to be published.
[26] T. A. Mohammed, S. Alhayali, O. Bayat, and O. N. Uçan, "Feature reduction based on hybrid efficient weighted gene genetic algorithms with artificial neural network for machine learning problems in the big data," *Sci. Program.*, vol. 2018, pp. 1–10, Oct. 2018.
[27] T. Weilkiens, J. G. Lamm, S. Roth, and M. Walker, "B: The V-Model," in *Model-Based System Architecture*. Hoboken, NJ, USA: Wiley, 2015, pp. 343–352.
[28] M. Al-Zewairi, M. Biltawi, W. Etaiwi, and A. Shaout, "Agile software development methodologies: Survey of surveys," *J. Comput. Commun.*, vol. 05, no. 05, pp. 74–97, 2017.
[29] Y. Zhou, "Predictive big data analytics using the UK Biobank data," *Sci. Rep.*, vol. 9, no. 1, p. 6012, Dec. 2019.

[30] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe, "Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease," *PLoS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0202344.

[31] W. Pearson, C. T. Tran, M. Zhang, and B. Xue, "Multi-round random subspace feature selection for incomplete gene expression data," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2019, pp. 2544–2551.

[32] S. Al-Janabi and A. F. Alkaim, "A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation," *Soft Comput.*, vol. 24, no. 1, pp. 555–569, Jan. 2020.

[33] J. Salvador-Meneses, Z. Ruiz-Chavez, and J. Garcia-Rodriguez, "Compressed kNN: K-nearest neighbors with data compression," *Entropy*, vol. 21, no. 3, p. 234, Mar. 2019.

[34] D. A. Clifton, J. Gibbons, J. Davies, and L. Tarassenko, "Machine learning and software engineering in health informatics," in *Proc. 1st Int. Workshop Realizing AI Synergies Softw. Eng. (RAISE)*, Jun. 2012, pp. 37–41.

[35] E. Boonchieng and K. Duangchaemkarn, "Digital disease detection: Application of machine learning in community health informatics," in *Proc. 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2016.

[36] J. Frochte and J. Frochte, "Python, NumPy, SciPy und Matplotlib—In a nutshell," in *Maschinelles Lernen*. Munich, Germany: Carl Hanser Verlag GmbH, 2019, pp. 32–67.

[37] W. Jentner and D. A. Keim, "Visualization and visual analytic techniques for patterns," in *High-Utility Pattern Mining*. Cham, Switzerland: Springer, 2019, pp. 303–337.

[38] M. I. Azeem, F. Palomba, L. Shi, and Q. Wang, "Machine learning techniques for code smell detection: A systematic literature review and meta-analysis," *Inf. Softw. Technol.*, vol. 108, pp. 115–138, Apr. 2019.

MOHAMMED MOREB was born in Hebron, Palestine, in 1981. He received the B.Sc. degree in information technology from Palestine Polytechnic University, and the M.Sc. degree in computer science from Al-Quds University. He is currently pursuing the Ph.D. degree in electronic and computer engineering with Altinbas University.

The focus of his research is software engineering in health informatics. He has over twelve years of experience in managing software development projects, including large government IT systems.

TAREQ ABED MOHAMMED received the B.Sc. degree in computer science from the College of Science, Kirkuk University, Kirkuk, Iraq, in 2007, the M.Sc. degree from Cankaya University, Ankara, Turkey, in 2012, and the Ph.D. degree in electronic and computer engineering from Altinbas University, Istanbul, Turkey, in 2019.

In 2019, he started teaching at the College of Computer Science and Information Technology, University of Kirkuk. He has advised many studies for M.Sc. and Ph.D. students at various universities, participated in many international conferences and contributed to various scientific studies.

OGUZ BAYAT received the B.S. degree from Istanbul Technical University, Istanbul, Turkey, in 2000, the M.S degree from the University of Hartford, CT, USA, in 2002, and the Ph.D. degree from Northeastern University, Boston, MA, USA, in 2006, all in electrical engineering.

He completed the Executive Certificate Program in Technical Management and Leadership at Massachusetts Institute of Technology, Boston, MA, USA, in 2009. Since 2011, he has been serving as a Professor with the Department of Electrical and Electronics Engineering, Altinbas University. He is also an Advisor to the President and the Director of the Graduate School of Science and Engineering, Altinbas University.

OGUZ ATA received the BSc degree Computer Engineering from Sakarya University in 2004, and the MSc degrees Computer Engineering from Beykent University in 2008, and PhD degrees in software engineering from Trakya Üniversitesi in 2012. He has been the head of Department at Software Engineering at Altinbas University and lecturer at Altinbas University. His research interests include software repository mining, software measurement and testing, process improvement, and requirements engineering