Machine Learning Methodology with Software Engineering in Health System: A Review on SEMLHI

Anjana Yasodharan[1], Elda Dani[2], Neetha Thomas[3]
[1]BCA Scholar
Santhigiri College of Computer Sciences,
Vazhitala, Thodupuzha, Idukki
bcaa19_2218@santhigiricollege.com

[2]BCA Scholar
Santhigiri College of Computer Sciences,
Vazhitala, Thodupuzha, Idukki
bcaa19_2246@santhigiricollege.com

[3]Assistant Professor
Department of Computer Science
Santhigiri College of Computer Sciences,
Vazhithala, Thodupuzha, Idukki
neethathomas@santhigiricollege.com

Abstract:Health care is a field in which the discipline of software engineering and machine learning necessarily co-exist. Software engineering is a detailed study of engineering to the design, development and maintenance of software. Machine Learning (ML) is a rapidly maturing branch of computer science since it can store data on a large scale. Therefore, this review is purely based on the interaction between software engineering and machine learning within the context of health system. In this review we introduce a novel framework for health informatics called the framework and methodology of SEMLHI (The Software Engineering for Machine Learning in Health Informatics). The SEMLHI framework supports the methodological approach to conducting research on health informatics. It also supports a structure that presents a common set of machine learning terminology to use, compare, measure, and design software system in area of health. Throughout this paper, discussing about the four modules of SEMLHI framework that organize the task in the framework using SEMLHI methodology. The four modules are: Software, Machine Learning Model, Machine Learning Algorithmand Health Informatics Data.

Keywords: Health Informatics, Methodology, Framework, SEMLHI, Machine Learning, Software Engineering.

1. Introduction

Software engineering is the systematic application of engineering approaches to the development of software. Machine learning (MI) is a rapidly maturing branch of computer science since it can store data on large scale. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.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.SEMLHI stands for The Software Engineering for Machine Learning in Health Informatics.The SEMLHI methodology is used in software development in the health area.The SEMLHI framework supports the methodological approach to conducting research on health informatics. The SEMLHI framework includes a theoretical framework to support research and design activities that incorporate existing knowledge. The SEMLHI framework was composed offour components that help developers observe the healthapplication flow from the main module to submodules to runand validate specific tasks. This enables multiple developersto work on different modules of the application simultaneously.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.

2. An Analysis Study On Health Industry

In the current tech landscape, dynamics of both hardware and software is changing.For instance, according to the UK Center for Health Solution report, 48 percent of medical devices are connected through IoT, which is expected to rise up to 68 percent in the next five years.Even the software applications used in hospitals, such as Appointment Management System, Patient Administration System, and Laboratory Information Management System are now getting powered by advance techs like AI and machine learning.To harness the potential of healthcare technology to transform the health systems and develop a connected healthcare environment, healthcare leaders and clinicians need to forge closer ties with medical manufacturers and software application development companies. They can share information and develop new business models and scenarios that can improve the adoption of new technology in healthcare.

3. SEMLHI

This review is based on the original data collected by the author Mohammed Moreb from a hospital run by the Palestine government. Then the remaining data were analyzed using the developed framework to compare ML techniques that predict test laboratory results. The proposed

module was compared with three system engineering methods, Vee, Agile and SEMLHI. Theresults were used to implement the prototype system, which requires a machine learning algorithm.

	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 high	slow	Very quick
Delivery	One-shot	One-shot	Incremental

Table 1: Comparison of three system engineering methods Vee, Agile and SEMLH

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 data collection was conducted by the author Mohammed Morebover the last three years, and 458k patients were identifiedwith corresponding patient nos. Overall, for the PMCdataset,141k patients with 1.63% missing, a mean of 1.08M, a std devof 554k, a min of 10k, and a max of 1.04M were included.For the age label, 141k patients with 1.63% missing, a meanof 32.24, a std dev of 26.25, a min of 0, a max of 88, and amedian of 29 were considered.

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, feature selection problems using genetic algorithms, and new gene-weight mechanisms. 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.

The SEMLHI methodology is used in software developmentin the health area. Thedevelopmentprocess includes many methodologies, such asthe waterfall methodology, spiral methodology, and agilemethodology, which can be used to define and developthesoftware for traditional applications. The results of the comparisonbetween SEMLHI methodology, the Vee methodology, and theAgile methodology are illustrated in Table 1. Developersfollow manysequence steps, such as design (encode data,define outliersand clean the data), implement (verification and validation),maintain defined workflows, structure information, providesecurity and privacy, test theperformance, and then releasethe software applications for developing the HI system.Records in most datasets in HIare weakly structured and non-standardized. 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.

3.1. SEMLHI Framework

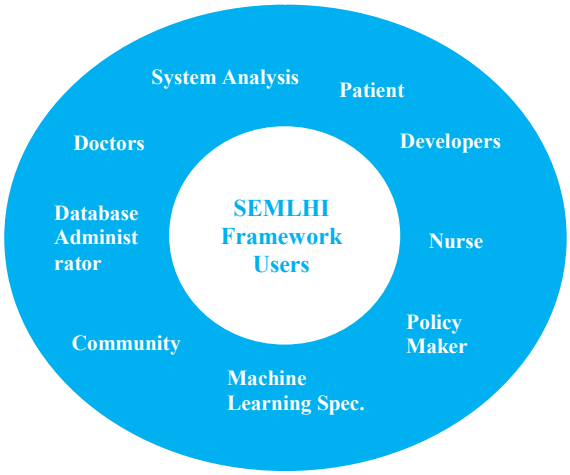

Figure 1: SEMLHI framework users

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,will work directly as developers or system analysts with approach frameworks or indirectly by using the results, as illustrated in Figure 1.Our framework was composed of four components or modules which are software, machine learning model, machine learning algorithms, and health informatics data. The below figure shows howeach module interacts with all modules to work as a framework.

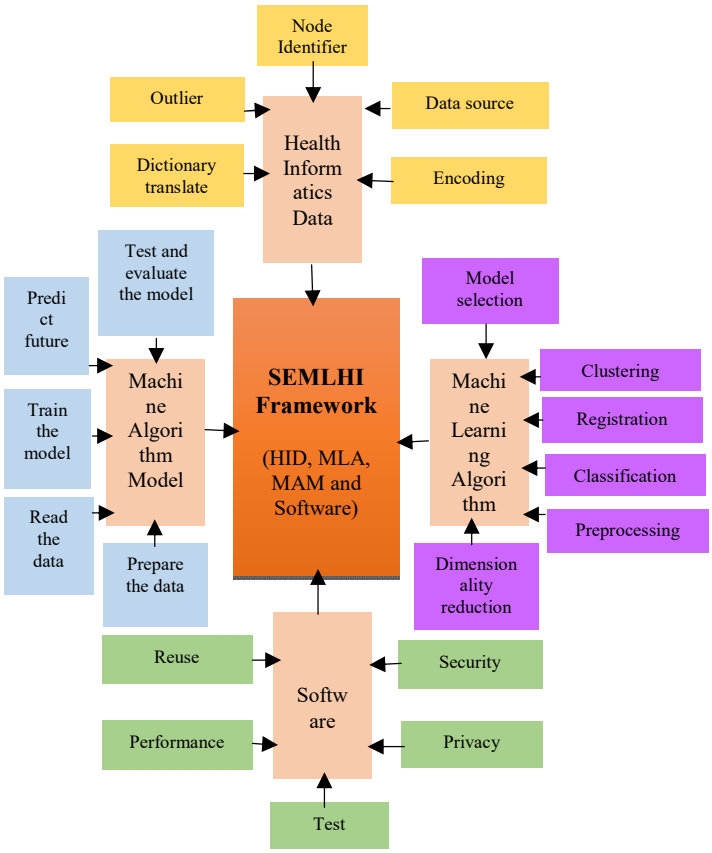

Figure 2:SEMLHI framework components

A. Health Informatics Data

In ML, data are essential, and choosing the methods for presenting and visualizing knowledge is the most important step.To use a dataset on health informatics 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. 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 foundand 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. To predict disease, they used ICD-10 with multiple labels, as each patient has an ICD
(International Statistical Classification of Diseases and Related Health Problems) code in their health records, which can affect all regions of the retina. However, there is currently no classification system for distinguishing anterior (peripheral) and posterior (macular) data.

B. Machine Algorithm Model

Machine learning helps us extract useful features from a dataset to address or predict health-related events. The machine algorithm model (MAM) component includes five submodules. They are, read the data, prepare the data, train the model, test and evaluate the model, and predict new data. Figure 4 describes the sequence of these stages. The main 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.

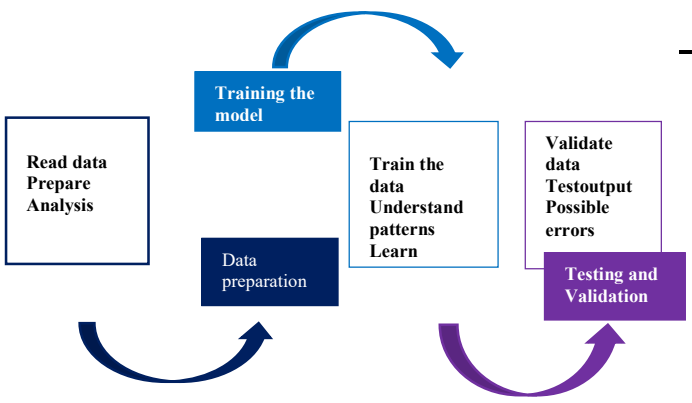

Figure 4: Mechanism of machine algorithm model

C. ML Algorithms

Machine learning algorithms (MLAs) are used to compute the parameters that might define a model, optimize its network topology and improve the system convergence without losing information. MLAs including some submodules they are listed in Table 2. As a supervised learning method, k-nearest neighbours (KNN) 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 3discusses the most of the machine learning algorithms used forhealth classification.

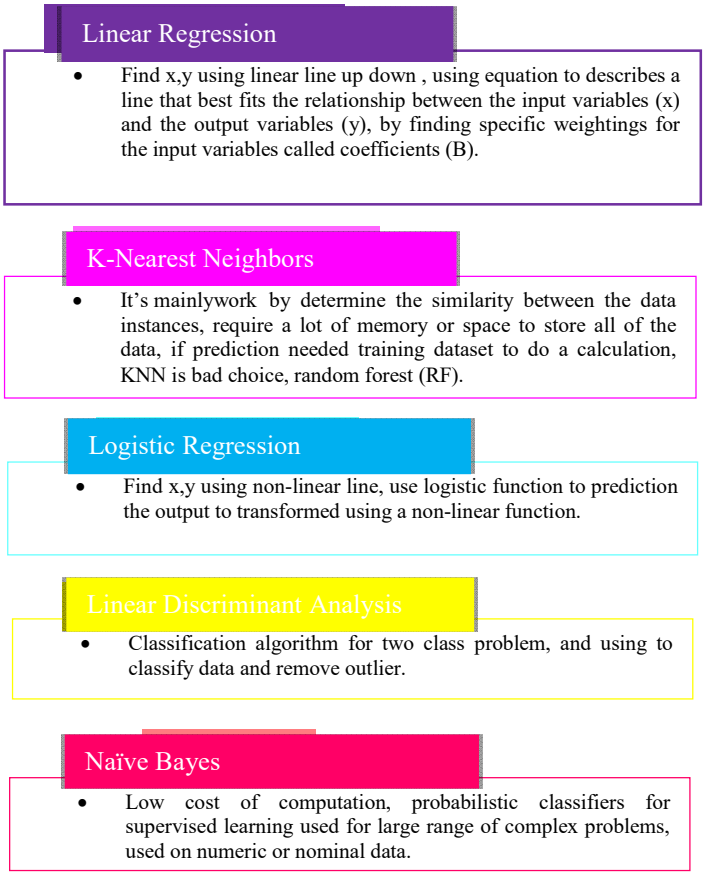

Figure 3: Machine learning algorithms used for health classification

SubModel for MLA Component	Applications	Algorithms
Classification	Spam detection, image recognition	Nearest neighbors, random forest, SVM
Re	Drug response, stock prices	SVR, ridge regression, lasso
	Customer division, grouping test outcomes	k-means, spectral clustering, mean-shift
	Visualization, increased efficiency	PCA, feature selection, clustering, non-negative matrix factorization
Model selection	Grid search,cross	

Preprocessing	validation, measurements Transforming input information	Future extraction.

Table 2: Machine learning algorithms sub models

D. Software

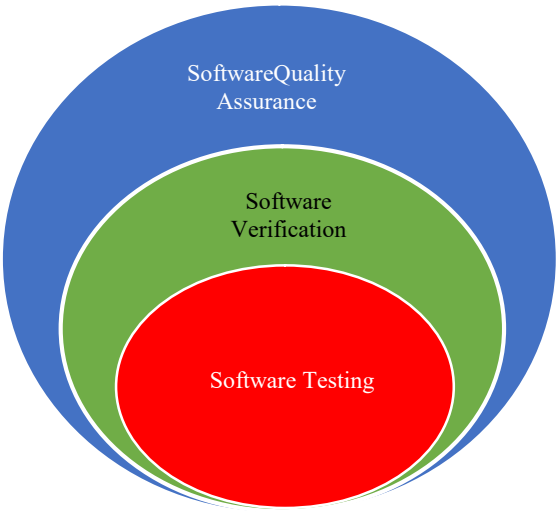

Figure 5: The software module includes subclasses including reuse, performance, testing, privacy and security

The software module,includes a subclass that includes reuse, performance, testing, privacy, and security as illustrated in Figure 5. 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. 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.

4. Conclusion

This paper introduced a new methodology, that can develop health informatics application using machine learning. The proposed methodology used the grounded theory methodology to develop SEMLHI framework. Developers use SEMLHI methodology to analyzing and developing software for the HI model and create a space in which SE and ML experts could work on the ML model lifecycle. SEMLHI methodology includes seven-phase, designing (encode data and Define outlier and cleaning up the data), implementing (Verification & Validation), maintaining and defined Workflows, structured Information, security and privacy, testing and performance, and reusing software applications. SEMLHI framework includes four modules that organize the tasks for each module, and introduce a SEMLHI Methodological that enable researchers and developer to analyze health informatics software from an engineering perspective. The ultimate goal from a SEMLHI Methodological is to define a standardized methodology for software development in the Health area and include all stages from defining the problem until developing the application.

We understood from this paper is that the SEMLHI framework includes four modules that organize the tasks for each module, and introduce a SEMLHI Methodological that enable researchers and developer to analyze health informatics software from an engineering perspective. And also understood that the ultimate goal from a SEMLHI Methodological is to define a standardized methodology for software development in the Health area and include all stages from defining the problem until developing the application.

References

[1] 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.

[2] M. Al-Zewairi, M. Biltawi, W. Etaiwi, and A. Shaout, ``Agile software development methodologies: Surveyof surveys," *J. Comput. Commun*,vol. 05, no. 05, pp. 74_97, 2017.

[3] T. A. Mohammed, S. Alhayli, S. Albawi, and A. Deniz Duru, ``Intelligentdatabase interface techniques using semantic coordination," in *Proc. 1st Int. Sci. Conf. Eng. Sci.-3rd Sci. Conf. Eng. Sci. (ISCES)*, Jan. 2018,pp. 13_17.

[4] 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.

[5] T. A. Mohammed, S. Alhayali, O. Bayat, and O. N. Uçan, ``Featurereduction based on hybrid efficient weighted gene genetic algorithms withartificial neural network for machine learning problems in the big data,"*Sci. Program.*, vol. 2018, pp. 1_10, Oct. 2018.

[6] J. Frochte and J. Frochte, ``Python, NumPy, SciPy und Matplotlib-In anutshell," in *MaschinellesLernen*. Munich, Germany: Carl Hanser VerlagGmbH, 2019, pp. 32_67.

[7] 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.

[8] J. Salvador-Meneses, Z. Ruiz-Chavez, and J. Garcia-Rodriguez, ``CompressedkNN: K-nearest neighbours with data compression," *Entropy*,vol. 21, no. 3, p. 234, Mar. 2019.

[9] F. Khomh, B. Adams, J. Cheng, M. Fokaefs, and G. Antoniol, ``Softwareengineering for machine-learning applications: The road ahead," *IEEESoftw.*, vol. 35, no. 5, pp. 81_84, Sep. 2018.

[10] Y. Zhou, ``Predictive big data analytics using the UK Biobank data," *Sci.Rep.*, vol. 9, no. 1, p. 6012, Dec. 2019.

[11] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, andN. M. Luscombe, ``Machine learning models in electronic health recordscan outperform conventional survival models for predicting patient mortalityin coronary artery disease," *PLoS ONE*, vol. 13, no. 8, Aug. 2018,Art. no. e0202344.

[12] Laplante,Philip(2007). What Every Engineer Should Know about Software Engineering. Boca Raton: CRC. ISBN 978-0-8493-7228-5. Retrieved 2011-01-21.

[13] ACM (2007). "Computing Degrees & Careers". ACM. Retrieved 2010-11-23.
[14] D. A. Clifton, J. Gibbons, J. Davies, and L. Tarassenko, ``Machine learningand software engineering in health informatics," in *Proc. 1st Int.WorkshopRealizing AI Synergies Softw. Eng. (RAISE)*, Jun. 2012, pp. 37_41.
[15] W. Jentner and D. A. Keim, ``Visualization and visual analytic techniquesfor patterns," in *High-Utility Pattern Mining*. Cham, Switzerland:Springer, 2019, pp. 303_337.

Author Profile

Anjana Yasodharanis born in 31st May 2001 in Muvattupuzha, Ernakulam, Kerala, India. She completed her higher secondary in SNDP HSS, Muvattupuzha and pursuing her under graduate in Bachelor of Computer Application in Santhigiri College, Vazhithala, Kerala.

Contact-bcaa19_2218@santhigiricollege.com

Elda Dani is born on 24th January 2001 in Vazhakkulam, Ernakulam, Kerala, India.She completed her higher secondary in St. Augustine's HSS, Kalloorkad and pursuing her under graduate in Bachelor of Computer Application in Santhigiri College, Vazhithala, Kerala.

Contact-bcaa19_2246@santhigiricollege.com

Neetha Thomas is an Assistant Professor of Computer Science Department in Santhigiri College of Computer Sciences, Vazhithala, Kerala. She is an MCA Qualifier and has 8 years of work experience in Santhigiri College, Vazhithala and NESS Technologies, Bengaluru. She won the award of Best Oral Presentation Award in 2019.

Contact- neethathomas@santhigiricollege.com