

Towards a disease prediction system: biobert-based medical profile representation

Rima Hatoum^{1,3}, Ali Alkhazraji¹, Zein Al Abidin Ibrahim^{1,3}, Houssein Dhayni², Ihab Sbeity^{1,4}

¹Computer Science Department, Faculty of Sciences, Lebanese University, Hadat Campus, Beirut, Lebanon

²Computer Science Department, Faculty of Sciences, Saint Joseph's University (USJ), Beirut, Lebanon

³CCE Department, Faculty of Engineering, Lebanese International University (LIU), Beirut, Lebanon

⁴Computer Science Department, Faculty of Sciences, Lebanese International University (LIU), Beirut, Lebanon

Article Info

Article history:

Received Jul 7, 2023

Revised Oct 27, 2023

Accepted Dec 2, 2023

Keywords:

Clustering

Coronary artery disease

Disease prediction

Healthcare

ABSTRACT

Predicting diseases in advance is crucial in healthcare, allowing for early intervention and potentially saving lives. Machine learning plays a pivotal role in healthcare advancements today. Various studies aim to predict diseases based on prior knowledge. However, a significant challenge lies in representing medical information for machine learning. Patient medical histories are often in an unreadable format, necessitating filtering and conversion into numerical data. Natural language processing (NLP) techniques have made this task more manageable. In this paper, we propose three medical information representations, two of which are based on bidirectional encoder representations from transformers for biomedical text mining (BioBERT), a state-of-the-art text representation technique in the biomedical field. We compare these representations to highlight the powerful advantages of BioBERT-based methods in disease prediction. We evaluate our approach efficiency using the medical information mart for intensive care-III (MIMIC-III) database, containing data from 46,520 patients. Our focus is on predicting coronary artery disease. The results demonstrate the effectiveness of our proposal. In summary, BioBERT, NLP techniques, and the MIMIC-III database are key components in our work, which significantly enhances disease prediction in healthcare.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Zein Al Abidin Ibrahim

Department of Computer Science, Faculty of Sciences, Lebanese University

Hadat Campus, Beirut, Lebanon

Email : zein.ibrahim@ul.edu.lb

1. INTRODUCTION

Researchers are actively developing modern methods to detect and prevent diseases, particularly chronic ones, at early stages to mitigate their impact on human lives. This involves merging machine learning with patient data from healthcare facilities like hospitals and clinics. This data, encompassing X-ray images, MRIs, textual information, lab results, and patient assessments, is crucial for accurate diagnosis through data analysis and machine learning techniques. However, initial data acquisition may have some randomness, necessitating refining and purifying steps to ensure its suitability for analysis and early disease detection.

Recently, there has been a significant convergence of biomedicine and data science, underpinning notable biomedical achievements. Artificial Intelligence and Machine Learning algorithms play a pivotal role in healthcare, particularly in disease prediction. Google's DeepMind Health initiative is a notable example, training software to detect over 50 eye diseases based on image features. In a comparison with diagnoses by eight doctors, the software's predictions achieved an impressive 94% accuracy rate [1].

In the realm of scientific research and biomedical practice, navigating the complexities of medical data presents significant challenges, particularly concerning availability and diversity of data types. A pivotal aspect lies in meticulously choosing appropriate data sources to bolster the efficacy of predictive models [2] and ensure alignment with practical expectations. Among the essential data sources are human language descriptions detailing disease diagnoses, treatment protocols, intuitive symptom reports, clinical annotations, prescriptions, and electronic medical records (EMR). These reservoirs of information are highly prized for their ability to unveil nuanced insights into a patient's health status, providing invaluable depth to medical analyses and decision-making processes.

Researchers face a challenge: unstructured, human language medical data poses difficulty for machines to interpret. To bridge this gap, informatics experts turn to machine learning techniques, particularly in natural language processing (NLP), ensuring a seamless transition from human language to machine-readable data. This issue has garnered attention from numerous researchers [3]-[17].

For instance, Zhang *et al.* [3] Proposed a disease prediction algorithm based on symptom similarity analysis, considering patients' intuitive symptom reports as pertinent features. The prediction is made by comparing these symptoms with those already associated with known diseases. In [4], authors employ the Stanford Parser and rely on word2vec to convert medical information notes into numerical representations for machine learning algorithms.

Additionally, Batbaatar *et al.* [5] present a health-related named entity recognition (HNER) method, using unstructured Twitter messages to predict named entities: diseases, symptoms, and drugs. They generate word-level features (based on word embedding and part-of-speech tagging) and character-level features (based on convolutional neural networks or CNN). These efforts demonstrate a concerted push to overcome the challenge of unstructured medical data for enhanced disease prediction.

Clinical nursing notes, containing subjective assessments and vital patient information, often lose detail when transferred to electronic medical records (EMRs), which many clinical decision support systems (CDSSs) heavily rely on. Gangavarapu *et al.* in [6], addresses the gap by leveraging unstructured nursing notes to develop CDSSs, employing fuzzy token-based similarity and deep neural architectures for disease prediction, resulting in improved performance compared to existing models. Their model is based on the term weighting and the word embedding (Doc2Vec) as a vector-space modeling.

In [7], Tsipouras *et al.* proposed a fuzzy rule-based system for the prediction of the coronary artery disease (CAD). A decision tree is constructed using a dataset containing 199 subjects, each represented by 19 features and other demographic, history data and lab examination. Rules are extracted from the decision tree and used to construct a fuzzy model. The accuracy of the system reached 73.4% when using fuzzy model while it was 58.3% when using the normal rules extracted from the decision tree.

Based on fuzziness and rough set theories, Setiawan *et al.* proposed in [8] a rule-based support system for CAD disease detection which they used as training dataset from California Irvine University and tested the model on data coming from several countries. As stated by the authors, the proposed system could provide coronary artery blocking better than angiography and cardiologist. The results were validated by three experts in cardiology.

The CAD detection problem was also addressed by Chen and Hengjinda in [9] in which they proposed an algorithm for predicting coronary artery disease (CAD) using a machine learning approach. They built a pooled area curve (PUC) and compared the results of two algorithms, support vector machines (SVM) and Naive Bayes. SVM showed higher accuracy than Naive Bayes in predicting CAD.

Wu *et al.* presented in [10] a deep learning methodology for extracting coronary artery centerlines from cardiac computed tomography angiography (CTA) images, which is relevant for coronary artery disease (CAD) diagnosis. This approach achieved high accuracy in detecting coronary arteries, offering potential assistance in CAD diagnosis. Krittanawong *et al.* further investigated the predictive ability of various machine learning algorithms for cardiovascular disease in [11], highlighting the promise shown by SVM and boosting algorithms. However, due to algorithm heterogeneity, no specific algorithm was deemed superior in their systematic review and meta-analysis.

Breast cancer was also among the diseases prediction problem that got interests in the field of healthcare. Magboo *et al.* focused in [12] on the classification of breast cancer recurrences in women. They compared four algorithms: Logistic regression (LR), Naive Bayes (NB), k-nearest neighbors (KNN), and SVM. The Wisconsin dataset, which is relatively small, was used for the analysis. LR was found to be the best algorithm based on the evaluation metrics used.

Another type of cancer was also addressed in the literature by Maulidina *et al.* in [13] aimed to classify hepatocellular carcinoma (HCC), a type of liver cancer, using machine learning techniques. They utilized particle swarm optimization (PSO) for feature selection and random forest (RF) for classification. The dataset included 192 patients, with 66 diagnosed with HCC. The PSO-RF method achieved 100% accuracy, precision, recall, and F1-score when five optimized features were selected.

Some works in the literature based on molecules to predict diseases like the work of Ouyang *et al.* in [14] aiming to predict associations between microRNAs (miRNAs) and diseases. They proposed the WeightTDAIGN framework, which integrates auxiliary information related to miRNAs and diseases. The model outperformed other benchmark models in accurately predicting miRNA-disease associations.

Starting from the same idea, Huang *et al.* in [15] developed the Metapath Associated Heterogeneous Neural Embedding (MEAHNE) model for predicting connections between miRNAs and diseases in heterogeneous networks. The model used deep learning techniques and a semantic-based attention mechanism to extract complex associated information from biological networks. MEAHNE outperformed existing models in terms of prediction accuracy.

Chen *et al.* developed in [16] metagenome-based human disease prediction by multi- information fusion and machine learning algorithms (MetaDR), a machine learning-based framework for predicting human diseases from microbiome data. The model addressed the limitations of previous methods by considering abundance profiles from both known and unknown microbial organisms and capturing the taxonomic relationship among microbial taxa. MetaDR achieved competitive prediction performance and discovered informative features with biological insights.

Grazioli *et al.* proposed in [17] The multimodal variational information bottlenecks (MVIB) Deep learning model for microbiome-based disease prediction. The model jointly analyzed gut microbial species-relative abundance and strain-level marker profiles. It achieved high performance on various disease cohorts, demonstrating its competitiveness and speed compared to existing methods.

The studies showcased here illustrate how machine learning and computational models are being applied across a spectrum of healthcare domains, from predicting CAD to classifying diseases and analyzing associations. They underscore the effectiveness of these approaches in advancing diagnosis, prediction, and understanding of various medical conditions. However, some studies also recognize limitations and suggest areas for further research and improvement. Readers can refer to [18]-[21] for more information about methods in the healthcare domain.

In this paper, we aim to tackle this challenge to build a machine learning-based model to predict implicit patient diseases. The Biomedical NLP technique elected for our research is the hottest released word embedding technique called, the BioBERT [22]. We specifically targets heart diseases, with a focus on coronary artery disease. The goal is to predict this disease using machine learning techniques applied to the MIMIC-III database, which contains information on 46,520 patients.

Our research is summarized by three orientations: first, conceptually wise, our data source is unstructured, especially we focus to exploit the information underlying within the patient disease history. Secondly, we propose in this work three types of representation of medical information: 1) Model 1 (M1): Raw-data based representation, 2) Model 2 (M2): Average BioBERT-based representation and 3) Model 3 (M3): BioBERT-based with clustering representation. Finally, several machine learning classifiers are applied to achieve the prediction task.

This paper is organized as follows: the prediction system is described in section 1. In section 2, we present the three representations methods cited above. Section 3 presents the comparative evaluation of the representations' performance and the results discussion. Finally, we conclude in section 4 and presents future works.

2. PROPOSED APPROACH

The proposed intelligent system is composed of three main entities:

- Data preparation entity: Its aim is to take the data and prepare the medical profiles of the patients.
- Feature extraction entity: In this entity, we extract numerical vectors from the medical profiles of patients to train the system during the training phase. This is the most challenging step.
- Training/Prediction entity: Here we train several classifiers per disease based on the features extracted from the medical profiles of patients to make later predictions.

In Figure 1 you'll see the overview of our intelligent healthcare prediction system (IHCPS). Consequently, we will now delve into detailing the various entities involved.

2.1. Data preparation

In our proposal, we focus on the medical history of the patient to predict the diseases. Usually, a patient visits a doctor who may diagnose a disease during the visit or not. Moreover, the patient may visit the same doctor or another doctor several times and more than one disease may be diagnosed. The first type of patients is the ones that have not been diagnosed as having the disease during all their visits. In this case, they are considered as negative patients, and their medical histories are composed of data coming from all their visits. The second type of patients is those diagnosed having the disease at least one time during their visits. In

this case, the patients are considered as positive ones, and their medical histories are composed of data coming from the visits before the first diagnosis of that disease.

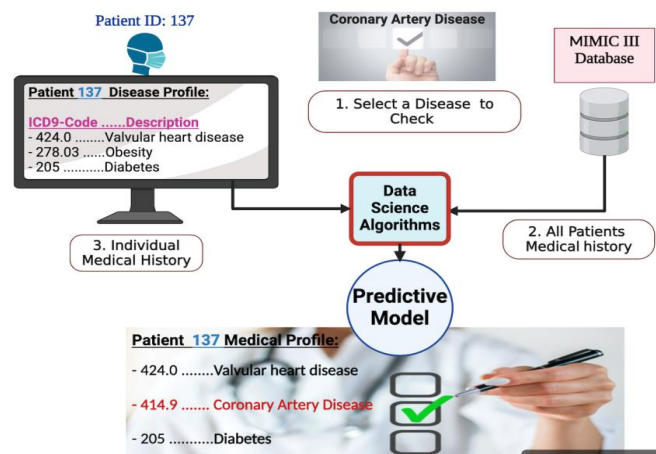


Figure 1. Intelligent healthcare prediction system

2.2. Features extraction

This step constitutes the core of our work; one of the main challenges of this domain is how to extract features from medical textual records of patients. We base on the latest pioneered text representation technique in the biomedical domain, the BioBERT. Three methods have been proposed and compared to extract features from healthcare information.

2.2.1. Raw-data-based representation

The main idea of this representation is to highlight all the diseases that have been diagnosed before the diagnosis of a specific disease. In other words, for each couple (D_i, P_j) in (diseases, patients), all the diseases that have been diagnosed before the diagnosis of D_i are highlighted. To highlight these diseases, a binary vector of size Naïve Bayes (NB) Diseases in which we put 0 at the position that corresponds to a disease that have not been diagnosed and 1 at the position corresponding to the disease that have been diagnosed before D_i for P_j . The process of features generation is explained in the Algorithm 1.

Algorithm 1. The steps and process of raw-data-based representation

Input: The input is the dataset to be used.

Step 1: Select a disease D_i from the set of diseases D .

Step 2: Select all patients that have been diagnosed with this disease after their first admission (after the first visit). Mark them as positive patients for D_i . The set will be noted P_i^+ . The target class here is the disease D_i .

Step 3: Mark the remaining patients as negative patients for D_i . The set will be noted P_i^- .

Step 4: For each positive patient $p_{i,j}^+$ in the set P_i^+ , highlight all the diseases D_k that have diagnosed before the diagnosis of D_i , generate a binary vector with the same size of the number of diseases in which you put 1 in all the cells corresponding to the highlighted diseases D_k and 0 in the remaining cells.

Step 5: Do the same process for negative patients in P_i^- .

Step 6: Balance the two classes in such a way the number of positive patients becomes equal to the number of negative ones. The balance is done by selecting random patients from the class having the more patients. Usually, the negative class contains more patients than the positive one.

Output: The output of the algorithm is a binary matrix noted binary representation matrix (BRM_i) for each disease D_i . The size of each matrix is $n*m$ where $n=2*\min(\text{number_positive_patients}, \text{number_negative_patients})$ and $m = \text{number_of_diseases} + 1$ (last column is the target value specifying if the patient is positive or negative for D_i). Figure 2 shows the characteristics of the BioBERT network.

2.2.2. Average BioBERT-based representation

With the advent of word embedding techniques that consider the contextual meaning of words and their customization for various domains, we've created word-embedding-based features to represent patients' previously diagnosed diseases. Specifically, we've employed BioBERT, a leading technique tailored for the healthcare domain, to generate 1,024-dimensional numerical vectors for each disease description. In our

approach, we construct a binary vector to highlight pre-diagnosed diseases before the diagnosis of a specific disease D_i for a patient $p_{i,j}$, followed by averaging all vectors representing the diagnosed diseases to generate features. The detailed process of feature generation is elucidated in the subsequent Algorithm 2.

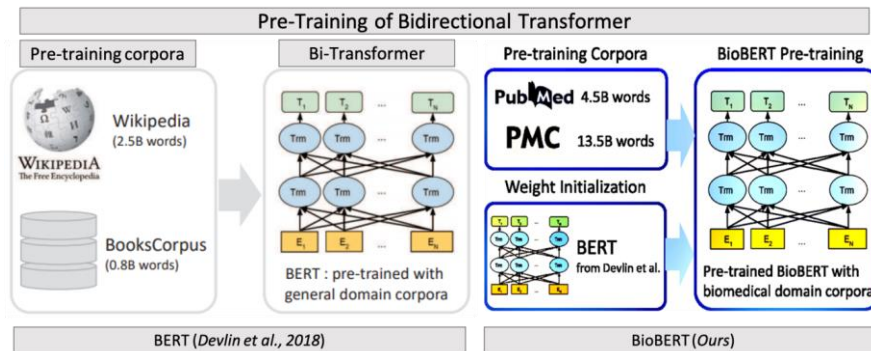


Figure 2. BERT and BioBERT representation

Algorithm 2. The steps and process of average BioBERT-based representation

Input: The input is the dataset previously presented.

Step1: Compute the BioBERT vectors for all the diseases based on their descriptions. For each disease D_i , its vector is noted V_i .

Step2: Select a disease D_i from the set of diseases D .

Step3: Select all patients that have been diagnosed with this disease after their first. Mark them as positive patients for D_i . The set will be noted P_i^+ . The target class here is the disease D_i .

Step4: Mark the remaining patients as negative patients for D_i . The set will be noted P_i^- .

Step5: For each positive patient $p_{i,j}^+$ in the set P_i^+ , highlight all the diseases D_k that have diagnosed before the diagnosis of D_i . Compute the average of all the vectors V_k of the diseases D_k .

Step6: Do the same process for negative patients.

Step7: Balance the two classes as presented in the raw-based representation.

Output: The output of the algorithm is a numerical matrix named Average BioBERT Representation Matrix (ABBRM) for each disease D_i . The size of each matrix is $n * 1,025$ where $n = 2 * \min(x,y)$ ($x = \text{number_positive_patients}$ and $y = \text{number_negative_patients}$) and $m = \text{size_of_BioBERT_vector} + 1$ (the last column is the target value highlighting if the patient is positive or negative).

2.2.3. Clustered BioBERT-based representation

With the help of BioBERT, we can reflect in our representation how many diseases are similar. To reduce the complexity of representation and to consider a smaller number of representatives of diseases, we propose to represent the very similar diseases by one vector. The idea comes from the healthcare domain where diseases are grouped into categories and each category contains sub-categories and so on. We do not delve into this hierarchy in our work, but we tried to group the diseases into groups where all the diseases in each group is represented by one vector, in our proposal, we have tested two clustering algorithms, k-means and hierarchical. We have chosen to use the Euclidean distance as measure of similarity between vectors.

As known, the k-means needs the number of clusters as initial input while the hierarchical one computes a dendrogram and we do not obtain clusters of diseases unless we cut the dendrogram to generate the clusters. That is why we have used optimal number of clusters calculation methods. For k-means, we have used the elbow to find the optimal (or sub-optimal) number of clusters while we have based on the dendrogram method of the SciPy library in Python to highlight it during the hierarchical process. During our experimentations, we have noted that the optimal number of clusters found matches the medical background knowledge about the similarity between diseases/descriptions. In Algorithm 3, we will explain the process of feature generation.

Algorithm 3. The steps and process of clustered BioBERT-based representation

Input: The input is the dataset previously presented.

Step1: Compute the BioBERT vectors for all the diseases based on their descriptions. For each disease D_i , its vector is noted V_i .

Step2: Apply a clustering algorithm to obtain M clusters. Let $C = \{C_k, k=1..M\}$ where M is the optimal number of clusters identified by the clustering algorithm. Each cluster may contain one or more BioBERT-based vectors.

Step3: Select a disease D_i from the set of diseases D .

Step4: Select all patients that have been diagnosed with this disease after their first admission (after the first visit). Mark them as positive patients for D_i . The set will be noted P_i^+ . The target class here is the disease D_i .

Step5: Mark the remaining patients as negative patients for D_i . The set will be noted P_i^- .

Step6: For each positive patient $p_{i,j}^+$ in the set P_i^+ , highlight all the diseases D_k that have been diagnosed before the diagnosis of D_i . Represent the patient by a binary vector $V_{i,j}$ of size M (number of clusters) where $V_{i,j}[k] = 1$ if the patient has one of the diseases in cluster C_k and 0 otherwise.

Step7: Do the same process for negative patients.

Step8: Balance the two classes in as presented in the raw-based representation.

Output: The output of the algorithm is a binary matrix Clustered BioBERT-based Representation Matrix (CBBRM_{*i*}) for each disease D_i . The size of each matrix is $n * M$ where $n = 2 * \min(\text{number_positive_patients}, \text{number_negative_patients})$ and $M = \text{number of clusters} + 1$ (the last column is the target value highlighting if the patient is positive or negative).

2.3. Training entity

After representing the disease history profile of patients by numerical values, it is now ready to create models for disease prediction. In our work, we have compared the performance of several classification methods with the different representations presented before to highlight the best couple (representation, classification method). In the next section, we present the experimentations done and the results obtained.

3. EXPERIMENTATIONS AND RESULTS

During experimentations, our focus was on utilizing the MIMIC III database [23], a vast and openly accessible repository comprising data from 46,520 actual patients treated at a medical facility, encompassing 623,369 records detailing diagnosed diseases during each admission. Our attention was particularly drawn to two specific types of data within the database, which we deemed pertinent for our research: 1) ADMISSIONS and 2) DIAGNOSIS-ICD records. The ADMISSIONS data provides all the information about the admissions of the patients to the medical center. The following information are provided: admission time, discharge time, patient ID, admission ID, diagnosis in the admission, location, language, and many other information. The DIAGNOSIS-ICD type focuses on the diagnosis of each patient during each visit. The disease is provided as an ICD9 code [24] while the textual and medical meaning of each disease ICD9 code is found in the D-ICD-DIAGNOSIS table containing all the ICD9 codes of diseases with their short-titles (abbreviated) and long-titles (full medical description).

In our study, we specifically focused on coronary artery disease (CAD), a leading cause of global mortality. Existing literature has relied on predefined features to detect CAD, but we believe there may be hidden features that could indicate its presence. Instead of defining specific features, we considered all diseases in the dataset, regardless of their direct relation to CAD. We used the MIMICIII dataset, consisting of 46,520 original testing patients, and employed cross-validation sampling with 10-folds for evaluation. We tested various classifiers, including KNN, SVM, RF, NN, NB, LR, AB, and SGD, and also created a meta-model from these models (excluding SVM). The evaluation was performed using the orange data mining software [25], and the results, including the optimal number of clusters (195) for k-means and hierarchical clustering, are summarized in Tables 1 to 5, considering different representations (M1, M2, M3), machine learning classifiers, and performance metrics.

Table 1. Statistics

Parameter	M1	M2	M3
Number of positive patients	881	881	881
Vector length per patient	6613	1024	195

According to Table 2, the best results obtained in term of accuracy and F1-measure is for the stacked classifiers using the second representation (accuracy= 79.6, F1-measure = 80.6). In Table 3, the precision and recall is the highest for the "Stack" using M2 representation. Table 3 and Table 4 show the performance according to the positive class only.

We can highlight in Table 4 and Table 5 that the classifiers perform slightly better for positive class than the negative class. This is important if a patient is predicted as having a disease, but it was not the real

case. This has less impact than predicting no disease for a patient while it was not the real case. In other words, the false positive is favorable than the false negative for this type of disease since more investigations and analysis will show that it is not the case.

Table 2. Average accuracy and F1-measure over the two classes + and -

Classifier	Accuracy				F1-Measure			
	M1	M2	M3 ¹	M3 ²	M1	M2	M3 ¹	M3 ²
KNN	68.8	75.2	66.1	67.9	68.8	74.8	64.7	67.0
SVM	49.8	71.9	58.3	63.0	39.4	71.5	54.3	61.1
RF	74.6	75.3	75.5	74.6	74.5	75.3	75.4	74.5
NN	74.2	78.7	73.4	73.6	74.2	78.7	73.4	73.6
NB	76.8	77.2	76.4	75.5	76.8	76.9	76.3	75.5
LR	76.6	79.1	75.3	77.0	76.6	79.1	75.3	77.0
AB	71.3	70.1	67.9	70.7	71.3	70.1	67.9	70.6
SGD	72.6	74.5	73.2	72.4	72.6	74.5	73.2	72.4
Stack	77.5	79.6	77.4	77.3	77.5	79.5	77.4	77.3

Table 3. Average precision and recall over the two classes + and -

Classifier	Precision				Recall			
	M1	M2	M3 ¹	M3 ²	M1	M2	M3 ¹	M3 ²
KNN	69.0	76.7	69.1	70.3	68.8	75.0	66.1	67.9
SVM	49.5	73.1	62.7	66.1	49.8	71.9	58.3	63.0
RF	75.0	74.7	75.8	74.9	74.6	74.7	75.5	74.6
NN	74.2	78.8	73.4	73.6	74.2	78.7	73.4	73.6
NB	76.9	79.1	76.7	75.8	76.8	77.2	76.4	75.5
LR	76.6	79.7	75.3	77.0	76.6	79.3	75.3	77.0
AB	71.4	70.1	67.9	70.7	71.3	70.1	67.9	70.7
SGD	72.6	75.0	73.3	72.4	72.6	75.0	73.2	72.4
Stack	77.6	79.2	77.5	77.3	77.5	78.8	77.4	77.3

Table 4. Accuracy and F1-measure of the + class

Classifier	Accuracy				F1-Measure			
	M1	M2	M3 ¹	M3 ²	M1	M2	M3 ¹	M3 ²
KNN	68.8	75.2	66.1	67.9	70.0	78.0	57.6	61.3
SVM	49.8	71.9	58.3	63.0	64.6	74.8	67.8	69.7
RF	74.6	75.3	75.5	74.6	76.1	75.4	76.8	75.9
NN	74.2	78.7	73.4	73.6	74.3	79.1	73.2	73.6
NB	76.8	77.2	76.4	75.5	76.6	79.8	77.6	76.7
LR	76.6	79.1	75.3	77.0	76.7	80.2	75.5	77.3
AB	71.3	70.1	67.9	70.7	71.7	70.1	67.7	70.0
SGD	72.6	74.5	73.2	72.4	72.9	74.5	73.9	72.7
Stack	77.5	79.6	77.4	77.3	78.2	80.6	78.2	77.7

Table 5. Precision and recall of the + class

Classifier	Precision				Recall			
	M1	M2	M3 ¹	M3 ²	M1	M2	M3 ¹	M3 ²
KNN	67.5	70.0	76.6	77.3	72.8	88.1	46.2	50.7
SVM	49.9	67.7	55.2	59.0	91.4	83.7	87.9	85.0
RF	71.9	75.2	72.9	72.1	80.9	75.5	81.0	80.1
NN	74.1	77.6	73.7	73.4	74.6	80.7	72.8	73.9
NB	77.5	71.7	73.8	73.2	75.6	89.9	81.7	80.6
LR	76.3	76.4	74.9	76.2	77.1	84.3	76.0	78.3
AB	70.8	70.0	68.0	71.6	72.6	70.3	67.4	68.6
SGD	72.0	74.4	72.1	72.0	73.9	74.7	75.7	73.4
Stack	76.0	76.8	75.7	76.3	80.5	84.7	80.8	79.2

4. CONCLUSION AND FUTURE WORKS

This work addresses disease prediction using machine learning based on patients' medical history. A key challenge was representing this history. Three methods were proposed and compared: 1) highlighting presence/absence of all diseases before a specific disease diagnosis, 2) using BioBERT to generate numerical vectors for diseases and averaging them for diagnosed diseases before a specific disease, and 3) clustering BioBERT vectors and highlighting presence/absence of a disease in each cluster. Various classification

methods were tested on generated features using the MIMICIII dataset (over 46,000 patients). Promising results were obtained for CAD prediction. Several future works are to be addressed: considering the order of disease presence, investigating other diseases (COVID-19, liver disease, diabetes), facilitating comparison with similar works, and incorporating additional information such as visit summaries, medication history, genetic/family information, lifestyle factors, and doctor/radiology/clinic staff reports. This could lead to a more accurate and efficient Medical Patient Profile (MPP) for disease prediction, patient grouping based on MPP similarity, and highlighting relationships between MPPs (e.g., family relationships).

ACKNOWLEDGMENTS





I extend my heartfelt appreciation to Lebanese University for providing essential resources. So, my sincere thanks to family and friends for their unwavering encouragement.

REFERENCES





- [1] J. D. Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342-1350, 2018, doi: 10.1038/s41591-018-0107-6.
- [2] A. L'Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine learning with big data: challenges and approaches," in *IEEE Access*, vol. 5, pp. 7776-7797, 2017, doi: 10.1109/ACCESS.2017.2696365.
- [3] P. Zhang, X. Huang and M. Li, "Disease prediction and early intervention system based on symptom similarity analysis," in *IEEE Access*, vol. 7, pp. 176484-176494, 2019, doi: 10.1109/ACCESS.2019.2957816.
- [4] S. Dubois, N. Romano, D. C. Kale, N. Shah, and K. Jung, "Learning effective representations from clinical notes," *arXiv: 1705.07025*, vol. 2, Jun 2017.
- [5] E. Batbaatar and K. H. Ryu, "Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach," *International Journal of Environmental Research and Public Health*, vol.16, no. 19, 2019, doi: 10.3390/ijerph16193628.
- [6] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. Kamath S., "Tags: Towards automated classification of unstructured clinical nursing notes," in *24th International Conference on Applications of Natural Language to Information Systems, NLDB*, 2019, doi: 10.1007/978-3-030-23281-8_16.
- [7] M. G. Tsiouras *et al.*, "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling," in *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 4, pp. 447-458, July 2008, doi: 10.1109/TITB.2007.907985.
- [8] N. M. Setiawan, P. A. Venkatachalam, and A. F. M. Hani, "Diagnosis of coronary artery disease using artificial intelligence based decision support system," in *In Proceedings of the International Conference on Man-Machine Systems*, Batu Feringhi, Penang, 2009, pp. 1-5.
- [9] J. L. Z. Chen and P. Hengjinda "Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study," *Journal of Artificial Intelligence and Capsule Networks*, vol. 3, no. 1, pp. 17-33, 2021, doi: 10.36548/jaicn.2021.1.002.
- [10] X. Wu, Y. Geng, X. Wang, J. Zhang, and L. Xia, "Continuous extraction of coronary artery centerline from cardiac CTA images using a regression-based method," *AIMS. Mathematical Biosciences and Engineering*, vol. 20, no. 3. pp. 4988-5003, 2023, doi: 10.3934/mbe.2023231.
- [11] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports*, vol. 10, 2020, doi: 10.1038/s41598-020-72685-1.
- [12] V. P. C. Magboo and M. S. A. Magboo, "Machine learning classifiers on breast cancer recurrences," in *Procedia Computer Science*, vol. 192, pp. 2742-2752, 2021, doi: 10.1016/j.procs.2021.09.044.
- [13] F. Maulidina, Z. Rustam, M. Novita, Q. S. Setiawan, and Sagiran, "Feature selection using particle swarm optimization and random forest for hepatocellular carcinoma (HCC) classification," *2021 International Conference on Decision Aid Sciences and Application (DASA)*, Sakheer, Bahrain, 2021, pp. 80-84, doi: 10.1109/DASA53625.2021.9682286.
- [14] D. Ouyang *et al.*, "Predicting multiple types of associations between miRNAs and diseases based on graph regularized weighted tensor decomposition," *Frontiers in Bioengineering and Biotechnology*, vol. 10, 2022, doi: 10.3389/fbioe.2022.911769.
- [15] C. Huang, K. Cen, Y. Zhang, B. Lio, Y. Wang, and J. Li, "MEAHNE: miRNA-disease association based on semantic information in heterogeneous network," *Life (Basel), MDPI*, vol. 12, no. 10, 2022, doi: 10.3390/life12101578.
- [16] X. Chen, Z. Zhu, W. Zhang, Y. Wang, F. Wang, J. Yang, and K.-C. Wong, "Human disease prediction from microbiome data by multiple feature fusion and deep learning," *iScience*, vol. 25, no. 4, 2022, doi: 10.1016/j.isci.2022.104081.
- [17] F. Grazioli, R. Siarheyev, I. Alqassem, A. Henschel, G. Pileggi, and A. Meiser, "Microbiome-based disease prediction with multimodal variational information bottlenecks," *Plos Computational Biology*, vol. 18, no. 4, 2022, doi: 10.1371/journal.pcbi.1010050.
- [18] Q. Mastoi, T.-Y. Wah, R. G. Raj, and U. Iqbal, "Automated diagnosis of coronary artery disease: a review and workflow," *Cardiology Research and Practice*, vol. 2018, 2018, doi: 10.1155/2018/2016282.
- [19] Md. S. Islam, Md. M. Hasan, X. Wang, H. D. Germack, and Md. N.-E-Alam, "A systematic review on healthcare analytics: application and theoretical perspective of data mining," in *Healthcare, MDPI*, 2018, doi:10.3390/healthcare6020054.
- [20] M. George and H. B. Anita, "Analysis of kidney ultrasound images using deep learning and machine learning techniques: a review," in *Pervasive Computing and Social Networking Conference*, vol. 317, Singapore, 2022, pp. 183-199, doi: 10.1007/978-981-16-5640-8_15.
- [21] S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Elsevier, Computers in Biology and Medicine*, vol. 156, 2023, doi:10.1016/j.compbiomed.2023.106668.
- [22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics, Journal of Leukocyte Biology*, vol. 36, no. 4, pp. 1234-1240, February 2020, doi: 10.1093/bioinformatics/btz682.
- [23] A. Johnson, T. Pollard, and R. Mark, *MIMIC-III Clinical Database CareVue subset (version 1.4)*, PhysioNet., 2016, doi: 10.13026/8a4q-w170.
- [24] *ICD9CM-International Classification of Diseases 9th Revision Clinical Modification*. USA: World Health Organization (WHO), 2010.
- [25] J. Demšar *et al.*, "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research*, vol. 14, pp. 2349-2353, 2013.

BIOGRAPHIES OF AUTHORS







Rima Hatoum     received the Ph.D. degree in Computer and Telecommunication Engineering from the Sorbonne University, Paris, France. She recently received her second Master degree in Information Systems and Data Intelligence from the Lebanese University, Beirut, Lebanon. She is currently an instructor in Lebanese International University, Lebanon. She has published several articles in conferences and journals sponsored by the IEEE, ACM. Her current research interests include machine learning healthcare applications, natural language processing, big data integration, and advanced telecommunication technologies. She can be contacted at email: rima.hatoum@hotmail.com.







Ali Alkhazraji     Holds a Master's degree in Computer and communications engineering from the Islamic University of Lebanon and he is currently pursuing a Ph.D. in the Informatics Department at the Lebanese University. His passion and dedication have led him to work on publishing research papers and present at international conferences, as he nears the completion of his Ph.D. He is poised to make significant contributions to academia, industry, and society as a whole. He can be contacted at email: ali.alkhazraji@ul.edu.lb.







Zein Al Abidin Ibrahim     Dr. Zein Al-Abidin Ibrahim is an associate professor at both the Lebanese University, Faculty of Science and at the Lebanese International University, Faculty of Engineering since 2012. He received his B.S. and master's degree in computer science from the Lebanese University, Faculty of Science in 2004 and his Ph.D. in Computer Science (Image, Information, Hypermedia) from the University of Paul Sabatier in Toulouse, France in 2007. Dr. Ibrahim worked as a research engineer at the Institute of Research in Toulouse (IRIT), France on the automatic and the hierarchical video classification for an interactive platform of enhanced digital Television. In 2008, he worked as a post doctorate at the INRIA-IRISA institute of research of Rennes for 16 months on TV stream structuring. Computer vision and machine learning including deep learning are among his research's topics of interests. He has several refereed journal and conference papers. Besides, he served as a reviewer for several conferences & journals. He can be contacted at email: zein.ibrahim@ul.edu.lb or zein.ibrahim@liu.edu.lb.



Houssein Dhayne     Ph.D. in Computer Engineering, specializing in Healthcare Data Integration, boasts over two decades of experience in software development. With seven years as a top-level manager, he co-founded and now serves as CEO at 3iSoft, a prominent IT company offering solutions and services for businesses of all sizes. His extensive background includes the development of large-scale information systems, such as Hospital Information Systems, Geographic Information Systems, and University Management Systems. With an academic background from Saint Joseph University, and a faculty position at USAL University, he brings a wealth of expertise to his work in a multidisciplinary and cross-disciplinary approach. He can be contacted at email: houssein.dhayne@3i-soft.com.



Ihab Sbeity     is a Professor in Computer Science at the Lebanese University. He received a *Maitrise* in applied mathematics from the Lebanese university, a Master in computer science - systems and communications from *Joseph Fourier university*, France, and a PhD from *Institut National Polytechnique de Grenoble*, France. His PhD works are related to Performance Evaluation and System Design. Actually, Dr. Sbeity occupies a full-time professor position at the Lebanese university – Faculty of Sciences I – computer sciences department. His research interests include software engineering, decision making, and deep learning applications. He can be contacted at email: ihab.sbeity@ul.edu.lb.