IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 14, No. 1, February 2025, pp. 385~396
ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i1.pp385-396

Machine learning for mental health: predicting transitions from addiction to illness

Ali Alkhazraji[1], Fatima Alsafi[1], Mohamed Dbouk[1], Zein Al Abidin Ibrahim[1, 2], Ihab Sbeity[1, 3]
[1]Department of Computer Science, Faculty of Sciences, Lebanese University, Hadat Campus, Beirut, Lebanon
[2]Department of Computer and Communications Engineering, Faculty of Engineering, Lebanese International University, Beirut, Lebanon
[3]Department Computer Science, Faculty of Sciences, Lebanese International University, Beirut, Lebanon

Article Info

Article history:

Received Mar 30, 2024
Revised Jul 27, 2024
Accepted Aug 30, 2024

Keywords:

Disease prediction
Machine learning
Medical profile
Mental illness
Word embedding

ABSTRACT

The increasing prevalence of infection-causing diseases due to environmental factors and lifestyle choices has strained the healthcare system, necessitating advanced techniques to save lives. Disease prediction plays a crucial role in identifying individuals at risk, enabling early treatment, and benefiting governments and health insurance providers. The collaboration between biomedicine and data science, particularly artificial intelligence and machine learning, has led to significant advancements in this field. However, researchers face challenges related to data availability and quality. Clinical and hospital data, crucial for accurate predictions, are often confidential and not freely accessible. Moreover, healthcare data is predominantly unstructured, requiring extensive cleaning, preprocessing, and labeling. This study aims to predict the likelihood of patients transitioning to mental illness by monitoring addiction conditions and constructing treatment protocols, with the goal of modifying these protocols accordingly. We focus on predicting such transformations to illuminate the underlying factors behind shifts in mental health. To achieve this objective, data from an Iraqi hospital has been collected and analyzed yielding promising results.

This is an open access article under the CC BY-SA 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

In today's digital era, technological advancements have brought about significant changes in lifestyles, contributing to the spread of diseases along with pollution. The healthcare sector is now faced with the urgent task of adapting to these challenges to ensure the well-being of individuals and the overall prosperity of nations. Rising healthcare costs and the increasing prevalence of physical and mental illnesses have made it more difficult for governments and health insurance providers to effectively address these issues. The digitization and processing of vast amounts of data have opened new possibilities for disease prediction, early intervention, and the development of innovative treatment options. Utilizing complex computation technologies, particularly artificial intelligence and machine learning algorithms, healthcare professionals can leverage various data sources such as images, text, and sounds to predict disease symptoms and take preventive measures well in advance.

The collaboration between the biomedical field and data science has led to significant advancements in disease prediction models and datasets over the past few years. Due to the global rise in addiction rates, researchers must seek solutions to address this prevalent issue. Additionally, addicts are undergoing treatment protocols and utilizing specialized systems for drug disposal. However, some individuals within this group might experience a shift in diagnosis from addiction to another psychological disorder. This change presents new challenges in implementing effective therapeutic protocols, as the underlying reasons for this transformation remain unknown. Hence, accurately predicting the progression from addiction to another psychiatric condition would enable skilled doctors to proactively implement preventive measures and provide researchers with the opportunity to investigate this phenomenon. The aim of this work is to address these ambiguous cases and provide insights into predicting and understanding the transformations that occur during the treatment of addicts, shedding light on the underlying factors behind these shifts in mental health. By leveraging the power of artificial intelligence and machine learning, this research endeavors to contribute to improved addiction treatment strategies and patient outcomes.

The remaining sections of the article will be organized as follows: Section 2 will be dedicated to previous work in the domain of disease prediction. In section 3, we will present an overview of word embedding (WE) methods used in NLP throughout this study. Section 4 will outline the proposed pipeline, while experiments will be presented in section 5, concluding in section 6 with proposed future extensions to the work.

2. PREVIOUS WORKS

In this section, since no previous works have addressed this disease yet, we will present a literature review of disease prediction in general. Literature can be organized in various ways. For example, we can categorize them based on the disease they address, such as heart attack, and corona virus, or based on the type of method used, such as traditional machine learning-based or deep learning-based methods. In our work, we will adopt the second type of organization.

2.1. Traditional machine-learning-based methods

A first type of methods in this category is the one proposed by Mall *et al.* [1] who developed a system to process X-ray images for bone fracture detection, employing the traditional machine learning methods. Initially, they analyzed and processed the images, emphasizing feature extraction using the gray level co-occurrence matrix (GLCM) method. Subsequently, they utilized a variety of machine learning methods, with the most optimal performance observed with the radial basis function (RBF) support vector machine (SVM) method.

Symons *et al* explores in [2]. The prediction of alcohol addiction treatment using a comparative study between clinical psychotherapists and machine learning. One notable finding in their work is that treatment is considered successful if the patient completes a minimum of 78 out of 84 days. While neural networks have shown significant progress, they were not utilized due to challenges in handling differently organized and unexplainable data as stated in their work. The researchers did not delve into exploring and working on this issue. Instead, they employed a traditional machine learning system, specifically random forest and logistic regression methods. Both methods yielded almost equal and high results, although the precise details of the data processing method were not elaborated upon in the article.

SVM and k-means were used by Sinha and Sharma [3]. More precisely, they used SVM as well as modified K-means, with some minor differences represented by adding an RBF kernel to work on selecting the good feature to use and yields to satisfactory results in the detection of coronary artery disease. At the time when modified K-Means proved to be effective in the work, since the data was non-linear, this result indicates the validity of this method with unstructured data. SVM, k-means, and 7 other classifiers were used in [4] in order to predict the coronary artery disease of patients based on three types of representations of medical history of patients with the help of BioBert. The results obtained showed the power of BioBert to represent medical text history of patients.

By analyzing an electroencephalogram (EEG) with the help of random forest classifier, Min *et al.* [5] presented a study to predict individual responses to electroconvulsive therapy (ECT) in schizophrenia patients. Using transfer entropy (TE) extracted from EEG data, a random forest classifier with feature selection accurately classified ECT responders and non-responders, achieving 85.3% balanced accuracy. Higher effective connectivity in frontal areas may indicate a favorable ECT response, offering potential for personalized ECT decisions in clinical practice.

Sleep apnea detection problem was addressed by Jezzini *et al*. [6]. In this study, automated methods for detecting sleep apnea using electrocardiogram (ECG), aiming to overcome the limitations of polysomnography (PSG). By comparing existing approaches and exploring various classifiers, the research evaluates the efficacy of artificial intelligence algorithms in sleep apnea detection. Results demonstrate that the K-nearest neighbors (KNN) classifier achieves a remarkable accuracy of 98.7%, surpassing other classifiers previously utilized in literature.

The KNN classifier, along with other classifiers, was evaluated for predicting the risk of developing coronary heart disease (CHD) in the next ten years in the study conducted by Minou *et al* [7]. Data preprocessing included cleaning the data by removing missing values, addressing imbalanced classes, and applying the synthetic minority oversampling technique (SMOTE) to prevent overfitting and data loss. Results indicated that the decision tree algorithm performed better than most classifiers, except for naive Bayes, KNN, and SVM, which also demonstrated acceptable performance.

KNN also showed its effectiveness to predict heart disease in [8]. The researchers analyze vast medical data to predict heart disease, utilizing supervised learning algorithms like naïve Bayes, decision tree, KNN, and random forest. Using a dataset from the Cleveland database comprising 303 instances and 76 attributes, this study focuses on 14 key attributes to assess algorithm performance. KNN demonstrates the highest accuracy score, offering insights into heart disease development probabilities.

Repaka *et al*. [9] developed a mobile application called smart heart disease prediction (SHDP) that uses machine learning techniques, specifically the naive Bayes classifier, to predict heart disease based on prior patient data. The application aims to identify risk factors for heart disease by categorizing user-submitted data using traditional machine learning methods. The team obtained basic data from UCI which was derived from information on previous patients. The evaluation of the system's accuracy rate takes less than 0.01 seconds and yields a result of 88.77% [9].

During the coronavirus disease 2019 (COVID-19) pandemic, where intensive care unit (ICU) resources are strained, a machine learning-based risk prioritization tool was developed to forecast ICU transfers within 24 hours, aiding frontline healthcare workers in efficient resource allocation and hospital flow management [10]. The work uses time series data from non-ICU COVID-19 admissions, a random forest model that was trained and evaluated on a retrospective cohort of 1987 patients. The proposed model demonstrated 72.8% sensitivity, 76.3% specificity, 76.2% accuracy, and 79.9% area under the receiver operating characteristics curve, offering a valuable screening tool for imminent ICU transfers and enhancing hospital resource allocation and patient throughput planning during the COVID-19 crisis.

Cheng *et al.* [10] conducted research to predict a patient's need to be admitted to the within 24 hours based on the health information stored in their health file. The dataset includes data for 1978 patients. Using the traditional machine learning approach and the random forest classifier, the evaluation results showed an accuracy rate of 67.2%. The aim of the research was to enable timely intervention and control of the patient's condition before it worsens.

A different type of machine learning method (multi-layer perceptron (MLP)) was used with other algorithms to predict covid-19 outbreaks. Ardabili *et al.* [11] explores the limitations of standard epidemiological models for predicting COVID-19 outbreaks and proposes machine learning and soft computing models as alternatives. Among various models examined, the MLP and adaptive network-based fuzzy inference system (ANFIS) show promise in predicting outbreaks. The study advocates for machine learning's effectiveness in modeling the complex and varied nature of COVID-19 outbreaks, suggesting its integration with traditional SEIR models for enhanced predictive capabilities.

Lastly, Priya and Jinny [12] presents another type of applications of the machine learning algorithms. This work evaluates machine-learning techniques, including support vector machines, gradient boosting, extreme gradient boosting, and random forest, to predict in vitro fertilization (IVF) pregnancy outcomes using Doppler and clinical parameters. Among these techniques, gradient boosting combined with random forest importance feature selection demonstrated the highest performance, achieving an accuracy of 82.3%. The results highlight the significance of ultrasound measurement parameters, particularly Doppler parameters, in influencing IVF outcomes.

2.2. Deep-learning-based methods

Deep-learning methods started to enter in the domain of healthcare especially for disease prediction. Zhang *et al.* [13] introduce a method for disease prediction and early intervention using a sentence similarity model. The approach incorporates WE and convolutional neural networks (CNN) to extract sentence vectors from patient data, focusing on symptom similarity analysis. The model integrates syntactic tree and neural network computations to enhance accuracy. Utilizing the Microsoft Research Paraphrase Identification (MSRP) dataset, the model achieved 83.9% F1 score and accuracy, demonstrating its effectiveness in predicting diseases and enabling early intervention. Additionally, experiments on the semantic textual similarity task showed promising results, indicating the model's capability to extract key information from sentences for disease prediction and early intervention.

CNN was also used in [14] which presents a method leveraging real-life international classification of diseases (ICD)-coded electronic medical records (EMR) from Africa (2018-2019) to develop a disease risk prediction model. Using CNN, the model processes EMR data into multi-step time-series forecasting data and predicts future disease risk based on demographic and medical history. Experimental results demonstrate an 80.73% accuracy in predicting individual disease risk. The model offers potential cost savings, pre-emptive corrective actions, and capacity planning insights for hospitals, providing peace of mind regarding personal health.

The power of deep-learning-based methods was highlighted in [15] in which a comparison of the results of applying traditional machine-learning-based and deep-learning-based algorithms for mortality estimations was conducted. The research conducted a comprehensive analysis of COVID-19 spread dynamics and mortality estimations in South Korea. Utilizing diverse data sets including demographic, location, and epidemiological information, the study revealed significant trends in pandemic progression. The analysis highlighted a peak in mortality rates during the initial and mid-phases of the outbreak, followed by a decline suggesting increased individual resilience. Moreover, the study emphasized the elevated risk among elderly individuals and males. Furthermore, the research employed various artificial intelligence-based models for predictive analysis of quarantined COVID-19 cases. Logistic regression, SVM, and deep learning sequential models achieved high accuracy rates, with the deep learning model yielding the highest accuracy of 99% through meticulous parameter tuning and prevention of overfitting. These findings provide valuable insights for healthcare practitioners to enhance their preparedness and response strategies.

The power of deep-learning-based algorithms has proven their efficiency also but this time in the prediction of acute kidney injury (AKI) [16]. This study aimed to compare the predictive accuracy of two models for AKI development in ICU patients. Utilizing urine output trends, a deep learning model was evaluated against logistic regression for AKI prediction. Results from 35,573 ICU patients revealed the deep learning model achieved higher accuracy (AUC=0.89, sensitivity=0.8, specificity=0.84) than logistic regression. Remarkably, the deep learning model anticipated 88% of AKI cases over 12 hours prior to onset, demonstrating its potential for early intervention and prevention in ICU settings.

When it comes to analysing medical images such as X-ray or CT images, here comes the power of deep CNN-based network as used in [17]. In this work, authors prove the utility of deep learning-based feature extraction frameworks for automatic COVID-19 classification, given the absence of clinically approved treatments. Various deep CNN, including MobileNet, DenseNet, and ResNet, were evaluated to extract features from chest X-ray and CT images. These features were then inputted into machine learning classifiers to differentiate COVID-19 cases from controls. Notably, the DenseNet121 feature extractor paired with a Bagging tree classifier demonstrated the highest performance, achieving 99% classification accuracy, followed closely by a hybrid ResNet50 feature extractor trained with LightGBM, achieving 98% accuracy. This approach holds promise for enhancing early detection and monitoring of COVID-19, potentially reducing mortality rates.

2.3. Natural language processing

Medical data and records contain various types of information, including textual reports and descriptions, which hold essential information for analysis. However, machine learning algorithms typically require numeric input data for processing. To bridge this gap, natural language processing (NLP) techniques offer a solution by transforming text into numerical vectors through embeddings techniques. These techniques convert textual information into high-dimensional vectors, enabling algorithms to interpret and analyze text-based data effectively. In the following, we will give an overview of embedding techniques in the literature:

- Word2Vec [18]: Word2Vec encodes the meaning of words into compact and dense vectors, called WE. These embeddings represent words in a continuous vector space, where words with similar meanings have vectors that are close together, often measured by cosine similarity. While Word2Vec models consider context during training by examining neighboring words, each word in the vocabulary is represented by a single vector. However, this approach may not fully capture all the nuances of word meanings and contexts.
- GloVe [19]: GloVe leverages the co-occurrence statistics of words in a corpus to capture global contextual information by constructing a word-word co-occurrence matrix based on the entire corpus. It exhibits notable performance in tasks such as word analogy and named entity identification. While GloVe competes with Word2Vec in certain tasks, its performance may surpass that of Word2Vec in others, underscoring its versatility and effectiveness in capturing semantic relationships in textual data.
- Embeddings from language models (ELMo) [20]: ELMo represents a different approach to WE compared to GloVe and Word2Vec. Instead of representing each word with a single vector regardless of its context, ELMo captures the context of the word within the entire sentence or phrase. It achieves this by producing different embeddings for the same word used in different contexts across various sentences. This contextualized word representation enables ELMo to capture nuances in meaning and syntactic structures more effectively.
- Bidirectional encoder representations from transformers (BERT) [21]: BERT is designed to develop models for specific tasks such as question-answering and language inference. It incorporates an additional output layer for fine-tuning, allowing for adaptation to different tasks and datasets. BERT is considered more straightforward and effective compared to traditional language modeling approaches due to its bidirectional architecture and contextualized embeddings.
- BioBERT [22]: BioBERT is a pre-trained model specifically designed for representing biological language in biomedical text mining applications. Unlike BERT, which is pre-trained on generic corpora such as Wikipedia and books, BioBERT is trained on biomedical corpora, including PubMed abstracts and PMC full-text articles. In a study by Lee *et al.* [22], BioBERT was fine-tuned on three biomedical text mining tasks: named entity recognition (BioNER), relation extraction (BioRE), and question answering (BioQA). The model's weights are initialized from BERT, resulting in notable performance improvements: BioNER saw a 0.62% increase in F1 score, BioRE improved by 2.80% in F1 score, and

BioQA achieved a 12.24% mean reciprocal rank (MRR) improvement. These results highlight BioBERT's effectiveness in various biomedical text mining tasks.

- Fusing BioBERT [23]: The model for BioNER integrates deep contextual-level WE, comprising an attention-based bidirectional long short-term memory (BiLSTM)-conditional random field (CRF) layer, a representation layer, and an input layer. By combining BioBERT, Word2Vec for WE, BiLSTM for character embedding, and contextual embedding (ELMo), notable improvements were observed across various datasets. Specifically, for datasets related to drugs and chemicals (BC5CDR-Chem and BC4CHEMD), the performance ratio increased by 0.7% and 0.95%, respectively. Similarly, for datasets linked to genes and proteins (JNLPBA and BC2GM), the model enhanced the performance ratio by 1.7%. In datasets associated with diseases (BC5CDR-disease and NCBI disease datasets), the model achieved an improved performance ratio of 2.0% and 2.1%, respectively.
- Clinical BERT [24]: Clinical BERT is employed to predict a 30-day hospital readmission across various admission times. Initially trained on clinical notes, BERT is then fine-tuned using the medical information mart for intensive care III (MIMIC-III) dataset, specifically focusing on hospital readmission. This model exhibits versatility beyond readmission prediction, proving useful in diagnosing conditions, estimating mortality risk, and assessing the duration of stay. Its efficacy extends to enhancing performance across multiple datasets, including diseases such as BC5CDR-disease and NCBI disease (2% and 2.1%, respectively), drug datasets like BC5CDR-Chem and BC4CHEMD (0.7% and 0.95%, respectively), and gene/protein datasets like JNLPBA and BC2GM (1.76% and 1.96%, respectively).
- BioAlbert [25]: The BioAlbert model employs two-parameter reduction techniques to address the challenges associated with scaling pre-trained models like BERT, which tend to have a large number of parameters, resulting in prolonged training times. BioAlbert, pre-trained on biomedical data, utilizes initial weights from Albert, leveraging PubMed, and PMC datasets. During the BioNER fine-tuning phase, the model demonstrates superior performance across various datasets, outperforming other approaches. Notably, in eight datasets, BioAlbert consistently delivers impressive results, surpassing competing models, as highlighted in the forthcoming comparison.
- BiLSTM + WE + character embedding (CE) [26]: A system devised for extracting chemical names from biomedical text integrates dynamic recurrent neural networks (RNNs), CRFs, and BiLSTM. Evaluation on the NCBI corpus, JNLPBA corpus, and BioCreative II GM corpus (gene) (disease) demonstrated notable performance, particularly excelling in the JNLPBA corpus for F1 score. However, BioBert and BiLSTM-CRF surpassed it in BioCreative II GM corpus and NCBI corpus, achieving scores of 89.98 and 90.84, respectively.
- Comparing F1 scores in the NCBI disease dataset among four models—BioBert, fusion of BioBERT, BiLSTM+WE+CE, and BioAlbert—BioAlbert achieved the highest result (97.18%) compared to the others. Traditional models mentioned earlier (Word2Vec, ELMo, and GloVe) were outperformed by BERT according to the state-of-the-art benchmarks mentioned. Additionally, ClinicalBert, applied to a distinct dataset, showed promising results with a 71.4% AUROC benchmark. Consequently, BioAlbert emerges as the top-performing model among those evaluated.

3. PROPOSED APPROACH

The objective of our study is to ascertain whether an addict may receive a distinct diagnosis based on their medical profile. We will work with two distinct organizations of the same dataset: one organized "by patients" and the other "by visits". In the former, all patient visits are aggregated into a single dataset without duplication. Consequently, each patient is represented by a single data file containing their profile, but not the details of individual visits. In contrast, the latter dataset retains all visit-specific information for each patient. The experimental process comprises four steps as shown in Figure 1.

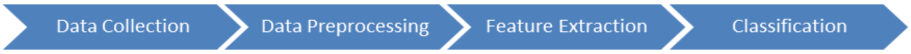


Figure 1. Proposed approach pipeline

3.1. Proposed approach pipeline

3.1.1. Data collection

We utilized authentic data sourced from the Ibn Roshd Hospital for Mental Illness and Addiction Treatment in Baghdad, Iraq. The dataset contains individual patient profiles along with their respective diagnoses. Three Excel sheets were collected:

- Main: This sheet encompasses the patient profiles.

– Sub_Table: It provides detailed information about each patient visit.
– Diagnosis: This sheet lists the diagnosis names and their corresponding codes.

a) Profiles

The patients’ profiles from aforementioned Iraqi hospital comprise a collection of personal and medical data structured as follows:

1. PAT_ID: Patient ID
2. GNDR: Gender
3. BRTH: Birth Date
4. VISIT_DATE: Date of each patient visit
5. LAB_TEST: Lab tests conducted by the patient during each visit
6. HSTRY: Patient's medical history
7. ECT: Indicates if the patient is undergoing electroshock therapy (0 or 1)
8. TREATMENT: Medication administered to the patient during each visit
9. DIAGNOSIS: Diagnosis assigned to the patient during each visit
10. CONDITION: Patient’s status during each visit

b) Diagnosis

Each diagnosis is accompanied by a certified code from the World Health Organization's diagnostic lists of diseases, known as (ICD9), which should be utilized on the main sheet. Furthermore, the diagnoses are documented in Arabic. The patients we are analyzing, who are addicts, are assigned codes ranging from F10 to F19. Table 1 presents the numerical information of the dataset while Table 2 lists some statistics about the dataset.

Table 1. Sample of data

ID	BRTH	GNDR	ECT	LAB	VST_DGNS_MDCN
2775	1994	F	0		<2010-12-04>F00<2010-12-19>G40<2011-02-13>G40<2021-07-04>F44
2777	1968	F	0		<2011-06-20>F32<2013-03-28>F51<2013-04-29>F51<2013-05-28>F51<2016-04-03>F34<2020-11-15>F41<2020-11-19>F41
2778	1955	M	0		<2011-08-14>F20 Olanzapine<2011-11-15>F20 Olanzapine<2012-06-26>F20 Olanzapine Citalopram (as HBr)<2013-09-17>F20 Olanzapine<2014-06-01>F20 Quetiapine<2014-10-14>F20Quetiapine<2015-01-22>F20Ecitalopram<2016-03-24>F20OlanzapineChlordiazepoxide<2020-0903>F20ChlordiazepoxideOlanzapine
2779	1979	M	0		<2011-08-14>F19 Fluoxetine HCl Chlordiazepoxide
2780	1982	M	0		<2011-08-14>F13Carbamazepine Citalopram (as HBr)

Table 2. Statistical information about dataset

Patients	9409	Patients Visits	57572
Addicts	1785	Addicts Visits	8080
Swapped Patients	104	Swapped Visits	1174

3.1.2. Data preprocessing

We employ data preprocessing, a data mining technique, to refine the raw data for further processing, ensuring it is presented in a meaningful and effective format. Given that the utilization of noisy data can lead to inferior findings despite employing robust algorithms, this preprocessing step is critical for achieving more precise results. The following measures will be implemented to achieve this goal.

a) Data cleaning

The following strategies are employed to address the challenges posed by missing and noisy data:

– Missing data: To manage missing data, we have two options: we can either exclude the entire tuple containing missing values, or we can replace the missing values with either the attribute mean or the most probable value. In our dataset, we have decided to discard the two columns labelled Lab (6392/8080) and Condition (7266/8080) due to the significant number of missing values they contain. For the remaining columns, any missing values will be substituted with 0.
– Noisy data: Machines are incapable of interpreting meaningless data, often referred to as noisy data. However, this aspect was not considered in our analysis.

b) Data transformation

This stage aims to transform the data into a format suitable for the mining process. The following steps will be undertaken to achieve this:

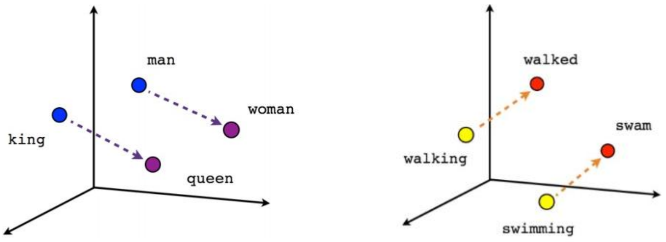
– Normalization: This involves scaling the data to a standard range. In our case, we will perform the following updates: i) Translate the "Age" column from gender (M, F) to numerical values (0, 1); and ii) Normalize the "Birth Date" column.

- Creation of new attributes: We will generate new attributes based on the provided list. For columns containing diverse information such as "history," "diagnostic," and "therapy," we will split the data into distinct columns, utilizing an approach similar to the one applied for other relevant columns.

These steps will ensure that the data is appropriately prepared for the subsequent mining process.

3.1.3. Feature extraction

The process of extracting essential information from raw data for diverse applications is called feature extraction. Given that most data possess numerical attributes, various methods exist to convert textual data into numerical vectors, aligning with the objectives of machine learning algorithms. Word vectorization, NLP technique, facilitates the conversion of words into vectors. Words are mapped onto vector spaces using a range of language models. Within this vector space, each word is depicted by a vector consisting of real numbers, such as ‘man’ and ‘woman’ for nouns or ‘walking’ and ‘walked’ for verbs, as shown in Figure 2.



Male-Female Verb tense

Figure 2. Similar word vectorization

Several models, particularly those tailored for the biomedical domain, are tasked with this responsibility, as discussed in the NLP section. To incorporate our data, we have opted for BioBert and BioAlbert. The three attributes utilized in this approach are history, treatment, and diagnosis. The following steps will be taken:

- Loading pre-trained models: We will load the pre-trained models for BioBert [27] and BioAlbert [28] using the hugging face application programming interface (API).
- Obtaining embeddings: We will utilize specific functions from BERT and ALBERT [28], [29] to obtain embeddings and characteristics.

It is worth noting that the feature counts extracted from the two models differ (see Table 3). Each feature is identified by its associated attribute name and unique number, such as "HSTR i," "TREATMENT i," and "DGNS i".

Hence, we now possess the finalized data ready for application in the subsequent phase, namely classification. The textual attributes have been replaced with their corresponding characteristics, and the categorical attribute has been encoded as (0.1). Additionally, the attribute "Age" has replaced the birth date.

Table 3. Number of extracted features

Model	Features
BioBert	768
BioAlBert	4096

3.1.4. Classification

Identifying and partitioning different data classes and concepts through a model is a crucial step in the data mining process, known as classification. The goal is to accurately predict the target class for each case based on the available data. This task was accomplished through the following steps, as depicted in Figure 3.

Steps:

- Ensure the data is prepared according to earlier phases.
- Select various classifiers.
- Employ the cross-validation approach to train and assess the models using 5-fold validation.

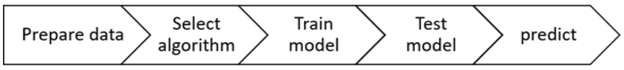


Figure 3. Classification actions

After we transformed textual data into numerical one using BioBERT and BioALBERT embedding techniques. we then utilized this data in two distinct ways. Firstly, by analyzing patient IDs regardless of the number of their visits and changes. Secondly, we considered the patient's frequent visits and the corresponding diagnosis at each visit. We analyzed the results using both balanced and unbalanced data, ignoring the date at all previous stages and focusing on the swap as the class feature. The two organizations of the data are as follows.

a) By patients

Under this version, two samples of data will be categorized: one with balanced classes and another with unbalanced classes. Both samples will incorporate features derived from BioBert and BioAlbert. It is important to highlight that the classifier "stack" integrates SVM, random forest, and neural networks.

i) Unbalanced classification

By using the two models of WE (BioBERT and BioALBERT), data when the target classes in a dataset are unenly distributed. Table 4 delineates the differing numbers of swapped and non-swapped patients in this scenario.

- BioBert: Table 5 reveals low precision and recall values for class 1. Specifically, the SVM model demonstrates a maximum precision of 0.83 but a relatively low recall of 0.36. Conversely, the Naive Bayes model achieves a maximum recall of 0.87, but its precision is approximately zero. Hence, the findings of this experiment may not be robust.

Table 4. Statistical information about unbalanced data

Models	Not Swapped	Swapped
BioBert	1600	82
BioAlbert	1668	99

Table 5. By patients-unbalanced-BioBert, BioAlbert results per class 1

Model	BioAlbert Accuracy (over the two classes)	Precision	Recall	Model	BioBert Accuracy (over the two classes)	Precision	Recall
AdaBoost	0.93	0.38	0.35	AdaBoost	0.95	0.48	0.54
Gradient Boosting	0.94	0.59	0.28	Gradient Boosting	0.96	0.61	0.53
Logistic Regression	0.94	0.55	0.29	Logistic Regression	0.96	0.66	0.51
Naive Bayes	0.48	0.08	0.87	Naive Bayes	0.52	0.08	0.87
Neural Network	0.94	0.44	0.26	Neural Network	0.96	0.68	0.51
Random Forest	0.94	0.61	0.22	Random Forest	0.96	0.67	0.53
SVM				SVM	0.96	0.83	0.36

Notably, high accuracy is observed for the "not swapped" class, which is expected given its majority representation in the dataset. Most models exhibit excellent outcomes, with precision and recall scores around 0.97 and 0.98, respectively.

- BioAlbert: In contrast, the results for BioAlbert models indicate relatively poor precision and recall values across the board. For instance, the random forest model achieves a maximum precision of 0.61 and a recall of 0.22. Similarly, naive Bayes achieves a maximum recall of 0.87 but minimal precision. Nevertheless, the inclusion of class 0 contributes to high accuracy.
- Comparison: Despite employing various classification strategies, both the BioBert and BioAlbert models yielded poor results overall. However, the BioBert model demonstrated slightly higher accuracy compared to the BioAlbert model, with a marginal difference of 0.02 in maximum accuracy between the two models.
- Discussion: The poor performance of both models can be attributed to the imbalance distribution of each class within the dataset. The algorithms demonstrate a tendency towards frequent class, with lower performance on the minority class. Essentially, the algorithms prioritize predicting the most frequent class, which leads to excellent overall results but at the expense of misclassifying minority instances. However, it's important to note that the misclassifications are not necessarily indicative of the algorithm weaknesses, but rather stem from the rarity of instances in the minority class.

In our experimentation, we initially utilized 7 models in the unbalanced classification setting. Subsequently, we augmented the number of models in the balanced classification setting.

ii) Balanced classification

In this experiment, the class distribution is equal (82 swapped and 82 not swapped patients), utilizing (BioBert and BioAlbert) with results in Table 6.

- BioBert: For class 1, the gradient boosting model achieves the highest accuracy of 0.85 and precision of 0.90. Although the recall (0.80) is slightly lower, it still compares favorably to other models, making this model the standout performer in the experiment. For class 0, gradient boosting also excels across all metrics with Accuracy, Precision, And Recall values of 0.85, 0.82, and 0.91 respectively.
- BioAlbert: For class 1, Gradient Boosting similarly demonstrates the highest accuracy (0.85) and precision (0.88), with recall (0.80) being close behind. Again, gradient boosting emerges as the top-performing model for this class. For class 0, gradient boosting achieves maximum values with accuracy, precision, and recall of 0.85, 0.82, and 0.89 respectively.

Table 6. By patents-balanced-Biobert, BioAlbert results per class1

	Biobert				BioAlbert		
Model	Accuracy	Precision	Recall	Model	Accuracy	Precision	Recall
AdaBoost	0.78	0.79	0.76	AdaBoost	0.72	0.75	0.67
KNN	0.52	0.52	0.53	KNN	0.49	0.49	0.50
Gradient Boosting	0.85	0.90	0.80	Gradient Boosting	0.85	0.88	0.80
Logistic Regression	0.81	0.81	0.81	Logistic Regression	0.81	0.82	0.78
Naive Bayes	0.72	0.67	0.85	Naive Bayes	0.71	0.67	0.82
Neural Network	0.82	0.84	0.80	Neural Network	0.73	0.72	0.76
Random Forest	0.84	0.87	0.81	Random Forest	0.79	0.81	0.75
SVM	0.79	0.80	0.79	SVM	0.74	0.73	0.77
Tree	0.78	0.79	0.78	Tree	0.74	0.76	0.70

Comparison: Both BioBert and BioAlbert models achieve their highest performance with the gradient boosting classifier. With a slight difference of 0.02 in precision for BioBert. However, other models show larger discrepancies, with differences of around 0.1.

Discussion: While the results in this phase are somewhat acceptable, they are not optimal. This can be attributed to the small size of the dataset. With only 82 patients for each class, which may not be sufficient for training models to achieve high accuracy.

Section discussion: Comparing precision and recall between the imbalanced and balanced classifications reveals a significant advantage for the balanced classification. In most models, the difference is substantial, except for SVM and naive Bayes, where the unbalanced classification outperforms in precision and recall respectively. This discrepancy is primarily due to the substantial class imbalance in the dataset, which impacts the accuracy of the models, as discussed previously.

b) By visits

Here, in this experiment, we utilized the second derived dataset organized by visits. Classification was conducted solely on balanced data, incorporating features extracted from BioBert and BioAlbert. The number of visits in each file is depicted in Table 7.

- BioBert: In Table 8, for class 1, gradient boosting outperforms all other models by achieving maximum values across all metrics: 0.98 in accuracy, precision, and recall. Similarly, for class 0, gradient boosting attains the best results with maximum values across all metrics: 0.98 for all.
- BioAlbert: According to the results in Table 8, for class 1, gradient boosting surpasses all other models by achieving maximum values in all metrics: 0.99 in accuracy and precision, and 0.98 in recall. Likewise, for class 0, gradient boosting attains the highest values: 0.99 in accuracy and recall, and 0.98 in precision.

Table 7. Statistical information about by visits data

Models	Visits
BioBert	2300
Bioalbert	2298

Table 8. By visits-balanced-BioBert, BioAlbert results per class1

	BioBert				BioAlbert		
Model	Accuracy	Precision	Recall	Model	Accuracy	Precision	Recall
AdaBoost	0.94	0.91	0.97	AdaBoost	0.94	0.91	0.97
KNN	0.86	0.82	0.93	KNN	0.86	0.81	0.94
Gradient boosting	0.98	0.98	0.97	Gradient boosting	0.99	0.99	0.98
Logistic regression	0.92	0.91	0.94	Logistic regression	0.83	0.86	0.79
Naive Bayes	0.72	0.71	0.73	Naive Bayes	0.71	0.72	0.67
Neural network	0.95	0.95	0.96	Neural network	0.95	0.95	0.96
Random forest	0.96	0.96	0.95	Random forest	0.96	0.97	0.95
SVM	0.50	0.50	0.30	SVM	0.55	0.62	0.26
Tree	0.93	0.92	0.94	Tree	0.93	0.91	0.96

Comparison: In this experiment, two models namely BioAlbert and BioBert were tested to measure their accuracy. The results show that BioAlbert model slightly outperforms BioBert. With a marginal difference of 0.01 between their maximum accuracies.

Section Discussion: The achieved results are very promising. With gradient boosting attaining high accuracy scores close to 1. This success can be attributed to the balanced nature of the dataset and the wealth of patient information available.

4. DISCUSSION

We conducted experiments on datasets organized both "by Patients" and "by Visits," with variations in class balance. While the results for unbalanced data were poor due to the scarcity of class 1 instances, those for balanced data were acceptable but not optimal. The limited dataset size, with only 82 patients per class, may have hindered the models' ability to achieve higher accuracy. Subsequently, focusing on the details of each patient visit in the "by Visits" approach yielded notably high and nearly perfect results, surpassing the "by Patients" approach. This emphasizes the efficacy of considering visit-level details for predictive purposes.

Regarding the BioBert and BioAlbert models, both exhibited comparable performance, with BioAlbert slightly edging ahead in the "by Visits" approach. However, in previous experiments, BioBert had a slight advantage in several instances. Overall, the performance of the two models was approximately equal. Several classifiers were employed in the classification phase, with gradient boosting consistently achieving high scores across most experiments. This suggests that gradient boosting is the most efficient classifier among those utilized when X axis represents no. of patients and Y axis represents no. of results as shown in Figure 4.

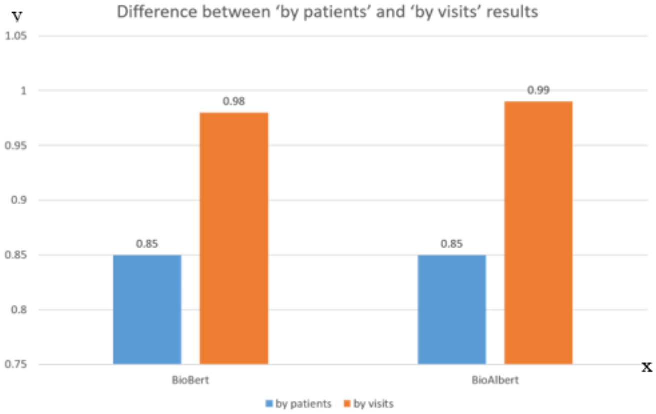


Figure 4. Final results

5. CONCLUSION AND FUTURE WORKS

Our research focused on disease prediction within the domain of addiction. We collected real data from a hospital in Iraq-Baghdad, comprising profiles of patients with various ailments. However, our study specifically targeted patients with addiction issues. The objective was to predict instances of relapse using various NLP models for text analysis and a range of classifiers for prediction. We divided the data into two samples: "by Patients" and "by Visits." The "by Patients" version lacked details regarding individual patient visits, unlike the latter. We explored both balanced and unbalanced datasets in the "by Patients" version, while exclusively focusing on balanced data for the "by Visits" dataset. Preprocessing stages were employed to prepare the datasets for text classification. We discussed different WE techniques and ultimately selected BioBert and BioAlbert for feature extraction. Subsequently, classifier models were trained using cross-validation technique (5-folds) for the prediction process. In the "by Patients" dataset, unbalanced data yielded unfavorable results, whereas the balanced dataset performed significantly better. However, when compared to the "by Visits" dataset, the latter exhibited superior performance. Both BioBert and BioAlbert models achieved similar results. Among the classifiers tested, gradient boosting emerged as the most efficient model. Ultimately, our proposed prediction model can aid medical professionals in anticipating the likelihood of an addict relapsing. This approach offers the potential to intervene and support individuals struggling with addiction, thus mitigating the risk of developing further illnesses. Numerous potential avenues for future research were identified, including experiments with different NLP models, considering patient visits as sequential data, refining the produced embeddings, and incorporating additional features. Our thesis lays the groundwork for further exploration in this area, providing valuable tools for analyzing diverse diseases using real-world datasets.

REFERENCES

[1] P. K. Mall, P. K. Singh, and D. Yadav, "GLCM based feature extraction and medical Xray," *2019 IEEE Conference on Information and Communication Technology,* Allahabad, India, 2019, pp. 1-6, doi: 10.1109/CICT48419.2019.9066263.

[2] M. Symons, G. F. X. Feeney, M. R. Gallagher, R. M. D. Young, and J. P. Connor, "Predicting alcohol dependence treatment outcomes: a prospective comparative study of clinical psychologists versus 'trained' machine learning models," *Addiction*, vol. 115, no. 11, pp. 2164–2175, 2020, doi: 10.1111/add.15038.

[3] D. Sinha and A. Sharma, "Automated detection of coronary artery disease using machine learning algorithm," *IOP Conference Series: Materials Science and Engineering*, vol. 1116, no. 1, 2021, doi: 10.1088/1757-899x/1116/1/012151.

[4] R. Hatoum, A. Alkhazraji, Z. A. A. Ibrahim, H. Dhayni, and I. Sbeity, "Towards a disease prediction system: biobert-based medical profile representation," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 2314–2322, 2024, doi: 10.11591/ijai.v13.i2.pp2314-2322.

[5] B. Min *et al.*, "Prediction of individual responses to electroconvulsive therapy in patients with schizophrenia: Machine learning analysis of resting-state electroencephalography," *Schizophrenia Research*, vol. 216, pp. 147–153, 2020, doi: 10.1016/j.schres.2019.12.012.

[6] A. Jezzini, M. Ayache, L. Elkhansa, and Z. Al A. Ibrahim, "ECG classification for sleep apnea detection," *2015 International Conference on Advances in Biomedical Engineering, ICABME 2015*, pp. 301–304, 2015, doi: 10.1109/ICABME.2015.7323312.

[7] J. Minou, J. Mantas, F. Malamateniou, and D. Kaitelidou, "Classification techniques for cardio-vascular diseases using supervised machine learning," *Medical Archives (Sarajevo, Bosnia and Herzegovina)*, vol. 74, no. 1, pp. 39–41, 2020, doi: 10.5455/medarh.2020.74.39-41.

[8] A. Golande and T. P. Kumar, "Heart disease prediction using machine learning techniques," *2023 International Conference on Artificial Intelligence and Smart Communication, AISC 2023*, vol. 8, no. 1S4, pp. 999–1005, 2023, doi: 10.1109/AISC56616.2023.10085584.

[9] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, vol. 2019, pp. 292–297, 2019, doi: 10.1109/icoei.2019.8862604.

[10] F. Y. Cheng *et al.*, "Using machine learning to predict ICU transfer in hospitalized COVID-19 patients," *Journal of Clinical Medicine*, vol. 9, no. 6, 2020, doi: 10.3390/jcm9061668.

[11] S. F. Ardabili *et al.*, "COVID-19 outbreak prediction with machine learning," *Algorithms*, vol. 13, no. 10, 2020, doi: 10.3390/a13100249.

[12] R. L. Priya and S. V. Jinny, "Elderly healthcare system for chronic ailments using machine learning techniques - A review," *Iraqi Journal of Science*, vol. 62, no. 9, pp. 3138–3151, 2021, doi: 10.24996/ijs.2021.62.9.29.

[13] P. Zhang, X. Huang, and M. Li, "Disease prediction and early intervention system based on symptom similarity analysis," *IEEE Access*, vol. 7, pp. 176484–176494, 2019, doi: 10.1109/ACCESS.2019.2957816.

[14] M. Krishnamoorthy, M. S. A. Hameed, T. Kopinski, and A. Schwung, "Disease prediction based on individual's medical history using CNN," *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*, pp. 89–94, 2021, doi: 10.1109/ICMLA52953.2021.00022.

[15] A. Sinha and M. Rathi, "COVID-19 prediction using AI analytics for South Korea," *Applied Intelligence*, vol. 51, no. 12, pp. 8579–8597, 2021, doi: 10.1007/s10489-021-02352-z.

[16] F. Alfieri *et al.*, "A deep-learning model to continuously predict severe acute kidney injury based on urine output changes in critically ill patients," *Journal of Nephrology*, vol. 34, no. 6, pp. 1875–1886, 2021, doi: 10.1007/s40620-021-01046-6.

[17] S. H. Kassania, P. H. Kassanib, M. J. Wesolowskic, K. A. Schneidera, and R. Detersa, "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 3, pp. 867–879, 2021, doi: 10.1016/j.bbe.2021.05.013.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pp. 1-12, 2013.

[19] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543, 2014, doi: 10.3115/v1/d14-1162.

[20] M. E. Peters *et al.*, "Deep contextualized word representations," *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 2227–2237, 2018, doi: 10.18653/v1/n18-1202.

[21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.

[22] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[23] U. Naseem, K. Musial, P. Eklund, and M. Prasad, "Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding," *2020 International Joint Conference on Neural Networks (IJCNN),* Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206808.

[24] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: modeling clinical notes and predicting hospital readmission," *arXiv-Computer Science*, pp. 1-9, 2019.

[25] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "BioALBERT: a simple and effective pre-trained language model for biomedical named entity recognition," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2021, 2021, doi: 10.1109/IJCNN52387.2021.9533884.

[26] S. Gajendran, D. Manjula, and V. Sugumaran, "Character level and word level embedding with bidirectional LSTM – Dynamic recurrent neural network for biomedical named entity recognition from literature," *Journal of Biomedical Informatics*, vol. 112, 2020, doi: 10.1016/j.jbi.2020.103609.

[27] Korea University Data Mining and Information Systems Lab, "BioBERT base cased v1.2," *Hugging Face*, 2021. [Online]. Available: https:/huggingface.co/dmis-lab/biobert-base-cased-v1.2/commits/main

[28] S. Arowili, "Bio M-ALBERT xxlarge SQuAD2," *Hugging Face*, 2021. [Online]. Available: https://huggingface.co/sultan/BioM-ALBERT-xxlarge-SQuAD2

[29] C. McCormick and N. Ryan, "BERT word embeddings tutorial," *Chris McCormick - Machine Learning Tutorials and Insights*, 2019. [Online]. Available: https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/

BIOGRAPHIES OF AUTHORS

Ali Alkhazraji holds a Master's degree in Computer and Communications Engineering from the Islamic University of Lebanon. He is currently pursuing a Ph.D. in the Department of Informatics 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.

Fatima Alsafi holds a Master's degree in Information Systems and Data Intelligence from Lebanese University, Faculty of Science. Her good experience in programming allows her to bring out the required results of articles, in addition to working on theoretical parts. She is a freelancer working on several projects with team work. She can be contacted at email: fatima.b.alsafi@gmail.com.

Mohamed Dbouk is currently a fulltime Professor at the Lebanese University (Department of Computer Science, Hadath, -Beirut, Lebanon), he coordinates (lead funder of): a master-2 research degree “ISDI-Information Systems & Data Intelligence”, and a research lab “L'ARICoD: Lab of Advanced Research in Intelligent Computing and Data”. He is co-founder/co-chair, of the “BDCSIntell: International Conference on Big Data and Cyber-Security Intelligence”. He maintains scientific partnerships with several universities (French, Canadian, England, Switzerland). He received his Ph.D. (Geographic Information System GIS & Hypermedia) from “Paris-11, Orsay” university, France, 1997. He has numerous refereed international publications, and he is scientific committee member of several international conferences and journals. His research’s topics of interests include software engineering and health care information systems, data-warehousing, big-data intelligence and data-mining, ubiquitous computing and smart cities, and service and cloud/edge oriented computing. He supervises many Ph.D. theses and academic research projects. Finally, he has strong background and experience in: academic auditing and enhancement (designer & auditor of academic curriculum, UNDP project) and in business administration and management (for two mandates; director of the Faculty of Sciences, Lebanese University, Hadath-Beirut). He can be contacted at email: mdbouk@ul.edu.lb.

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. He worked as a research engineer at the IRIT institute of research in Toulouse, 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.

Ihab Sbeity is a Professor in Computer Science at the Lebanese University. He received a Maîtrise in Applied Mathematics from the Lebanese university, a Master in Computer Science - systems and communications from Joseph Fourier University, France, and a Ph.D. from Institut National Polytechnique de Grenoble, France. His Ph.D. works are related to performance evaluation and system design. Actually, he occupies a full-time professor position at the Department of Computer Sciences, Faculty of Sciences I, Lebanese University. His research interests include software engineering, decision making, and deep learning applications. He can be contacted at email: ihab.sbeity@ul.edu.lb.