

A Novel Deep Similarity Learning Approach to Electronic Health Records Data

VAGISHA GUPTA¹, SHELLY SACHDEVA¹, AND SUBHASH BHALLA², (Member, IEEE)

¹Department of Computer Science and Engineering, National Institute of Technology Delhi (NITD), Delhi 110040, India

²Department of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan


Corresponding author: Shelly Sachdeva (shellysachdeva@nitdelhi.ac.in)

ABSTRACT The past decade has seen a tremendous advancement in using Electronic Health Records (EHRs) to offer clinical decision support and provide personalized healthcare to patients. Despite the potential benefits offered by EHR data, it is challenging to represent and analyze large EHRs for predictive modeling due to heterogeneity, high dimensionality, and sparsity. This article proposes a novel supervised Deep Similarity Learning approach that learns the patient representations and also finds the relationship between the patients using pairwise similarity learning to facilitate predictive analysis for personalized healthcare. We develop CNN_Softmax which is a Siamese-based neural network for multi-class classification methods corresponding to the prediction of disease. It uses Convolutional Neural Network (CNN) to study the vector representation of raw EHRs and capture essential information of patient features, and a Softmax-based supervised classification method that learns the similarity between pairs of patients and performs disease prediction using this similarity information. Our approach uses data type mapping to handle heterogeneity and the polynomial interpolation method to handle sparsity existing in EHR data. ORBDA, which is an openEHR (standard) benchmark dataset, is used for evaluating this study. Experimental results show that CNN_Softmax achieves an accuracy of 97.8%, a recall of 98.1%, a precision of 96.02%, and an F1 score of 97.82%. The comparative results show that our proposed novel methodology performs disease prediction with highly promising results and outperforms state-of-the-art similarity learning methods. The current study is the first attempt to perform disease prediction on standardized EHRs, to the best of the authors' knowledge. The deep similarity learning approach provides support for clinical decision making that is more reliable and generalizable than previous approaches and focuses on dealing with heterogeneous and sparse data. The concept also serves as a new implementation of artificial intelligence technologies for the application of clinical big data.

INDEX TERMS Convolutional neural networks, deep learning, electronic health records, nephrology, similarity learning, softmax-based technique.

I. INTRODUCTION

Over the past decade, with the advancements in technology, medical data has generally been stored in digital form in the healthcare sector. The health of patients is a priority, and medical experts are continually trying to implement new technologies and achieve significant results. Using decision support, medical practitioners can discover knowledgeable insights that help provide better diagnosis and treatment to patients. Secondary use of Electronic Health Records (EHRs) makes it possible to analyze a large amount of medical data

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara .

that can be used in different areas of research such as clinical decision support, information extraction, phenotyping, disease inference, and personalized healthcare [1].

Predictive analytics is one of the critical fields of clinical science to offer improved care to patients. Predictive analytics helps to optimize the cost of resources and provide better services to the patient in an environment of limited resources. It helps to make the diagnosis faster and helps doctors provide the patient with speedier and better treatment. It is a method of translating current data into new, unexpected situations and environments that can contribute to a more profound knowledge of disease-related information, such as better monitoring of different phases of illness, and identifying early onset

of illness. The result is improved quality of service to clients, increased support for physicians, and easy validation of standard diagnostic procedures for disease, and the provision of personalized healthcare to patients. In the last few years, most of the methods used to evaluate rich EHR data for predictive models were based on standard machine learning and statistical techniques which are not commonly and reliably used in making precise decisions about a complicated problem such as the outcome of clinical trials [2]–[4].

The performance of predictive algorithms depends primarily on data representation and feature selection [5], [6]. However, it is challenging to represent and analyze large scale EHR data due to high dimensionality, heterogeneity, and sparsity and systematic bias [7], [8]. These challenges have made it hard to detect the patterns that generate predictive models for clinical decision support. It is widely held that most of the effort in statistical models is hand-crafting, analyzing, choosing, and evaluating raw EHR data which can be time-consuming and often requires trial and error. Another challenge is that the possible predictive variables in EHR can potentially range in the thousands, especially when free clinical notes from physicians, nurses, and other service providers are used. Traditional predictive models deal with this challenge by simply choosing frequently occurring variables. This method can be troublesome, as the resulting models may deliver imprecise predictions that can overwhelm doctors [7]. However, recent advancements in deep learning and artificial intelligence may allow us to deal with these challenges. Deep learning techniques [9] can be applied to learn and extract the optimal features without any human input, allowing automatic detection of hidden patterns and relationships between features while building classifiers or predictors. Recently, in deep learning, various patient representation learning methods, such as recurrent neural networks (RNNs), autoencoder, deep reinforcement learning, and convolutional neural networks (CNNs) [5], [10], are being widely applied. Traditional deep learning approaches help us in finding out the essential characteristics of patients. However, deep learning approaches are not sufficient to train a model and cannot find the relationships between different patients and provide personalized treatments.

It has also become essential to find out relationships between patients. Evaluation of patient similarities is an important technique for many clinical research trials, such as risk stratification (clustering patients due to their vulnerability from a medical condition), comparative efficacy analysis, and predictive modeling. Similarity learning [11] of patients measures the similarity among a pair of patients based on their past medical data in a clinical encounter and has application in tasks related to classification and regression. Measuring similarity could contribute to building more accurate or efficient classifiers which will help in various applications like personalized medicines [12], trajectory analysis [13], drug discovery [12], and clinical trial patient recruitment [14]. Electronic Health Records (EHRs) can be used to evaluate clinical similarities between patients.

If physicians identify similar cases, the probability that the patient may be effectively treated will be significantly increased. Therefore, measuring the similarity between patients to provide personalized healthcare for applications such as patient-specific treatments, medications, and recommendations is an essential task.

This article proposes a novel deep similarity learning approach that we call “CNN_Softmax” for predictive analysis, by using patient similarity on standardized EHR data. Our approach consists of two parts: patient representation learning and patient similarity learning. For representation learning, we use Convolutional Neural Network (CNN) to capture essential information and regular patterns in the large-scale raw EHRs and obtain a vector representation containing important features of the original patient data. For similarity learning, we use a Softmax-based supervised classification technique, which is a Siamese-based neural network method [11] to perform multi-class classification on learned patient representations by measuring the similarity probability between similar pairs of patients and dissimilar pairs of patients. The similarity probability between a pair of patients is measured to indicate the risk of having the same disease between two patients. After obtaining the patient similarity information, we then perform disease prediction. The significant challenges [6] encountered while dealing with raw EHR and performing representation learning and similarity learning are heterogeneity and sparseness in data. We handle these challenges using suitable approaches, as illustrated in current research. In conclusion, we make several contributions as follows:

- We propose an end-to-end deep similarity learning approach that jointly learns the patient representations and finds the relationship between the patients using pairwise similarity probability. The approach could contribute to building more accurate and efficient classifiers which will help in providing personalized treatment to patients.
- We use Convolutional Neural Network (CNN) to study the vector representation of raw EHRs and capture essential information of patient features, and a Softmax-based supervised classification method to learn the similarity between pairs of patients.
- Our approach applies data type mapping to handle heterogeneity, and the polynomial interpolation method to handle sparsity, as these serve as significant challenges existing in large-scale EHR data.
- We are working on an openEHR compliant dataset, which is a standard for semantic interoperability. The main objective of the standardized data is to provide interoperability of EHRs among medical institutions.
- We perform experiments on standardized EHRs, and the results demonstrate that our proposed approach advances performance on predictive analysis tasks as compared to other state-of-the-art methods.

The organization of the paper is as follows: in Section II, we give a background study on deep learning, CNN,

and EHRs. Section III puts our work in a broader context by examining the state of the art in the current domain. Section IV illustrates the problem statement of our research. In section V, we describe the steps needed for disease prediction through a deep similarity learning-based approach. Section VI provides experimental results obtained after evaluating our model on a nephrology dataset, and section VII contains the discussion section.

II. BACKGROUND

This section briefly discusses important background concepts related to this research, including deep learning, convolutional neural networks, and electronic health records.

A. DEEP LEARNING

Deep learning has proven to be useful in many areas, including but not limited to image recognition, stock market prediction, automated text generation, and automated machine translation. Deep learning has also produced fruitful results in the healthcare domain. Drug discovery, precision medication, medical imaging, and disease prediction is being actively applied in this domain [15]. Deep learning is an artificial intelligence subdomain influenced by the structure of biological neurons connected in the brain. Deep learning helps to process the data and to identify hidden patterns for decision making. Deep learning models are multi-layered networks with the transformation between neurons in each layer. In the layered architecture of deep learning, upper layers extract the high-level features, and lower layers extract the lower level features.

B. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Network (CNN) is a deep neural network multi-layered architecture that has the most common use for visual tasks [16]. The three main advantages of using CNN as a deep learning architecture are finding high-dimensional sparse data, sharing of parameters, and identifying equivalent representation [17]. A convolutional neural network comprises an input layer, output layer, and many hidden layers. Hidden layers of the convolutional neural networks mostly consist of series of convolutional layers followed by subsampling layers and fully connected layers. An input layer receives specific inputs and produces a dot product with weights defined as certain filters. An activation function maps the input and output, and then a final convolution involves backpropagation to weigh the end product more accurately.

C. ELECTRONIC HEALTH RECORDS

An Electronic Health Record (EHR) is health information of an individual stored digitally and is instantly and securely available to authorized users. EHRs contain patient's diagnoses, medications, vital signs, treatment plans, progress notes, radiology images, and test results.

Classifications of records exist as unstructured and structured EHR data. Unstructured EHR data are written

or dictated notes based on a clinical context that describes a patient's condition and is most useful for clinical documentation. However, due to the unstructured format, ample typing and spelling errors, and the use of acronyms, abbreviations, and idiosyncrasies, they become challenging for computer analysis. Classification of structured EHR data may be as administrative data and ancillary data. Administrative data either remains unchanged throughout clinical encounters (such as demographic data) or keeps changing over time (such as diagnoses and procedures). Ancillary data is either discrete (such as physiological measurements, medications, and laboratory tests) or continuous (such as respiration and blood pressure). In previous research [18], work on structured and unstructured healthcare data was performed by proposing a CNN-based multimodel disease risk prediction algorithm. The current study is the first work focusing on both types of data in the field of big data analytics.

OpenEHR is an open standard that defines specifications in health informatics for storage, retrieval, and sharing of EHRs [19]. Recent pandemics (COVID-19, etc.) increase the need for using EHR from a research perspective. We use ORBDA (OpenEHR Benchmark Dataset), which is a standardized dataset for semantic interoperability of electronic health records provided by the Brazilian Public Health System (SUS) [20]. For analysis and disease prediction, we use a portion of ORBDA (Nephrology).

III. RELATED WORK

This section presents a review of some related literature on deep learning techniques and its applications for predicting diseases and building models with similarity learning techniques.

A. DEEP LEARNING

The use of Deep Learning has become widely popular and extensive in many applications, such as natural language processing, image classification, and fraud detection. In [15], [17], a general survey of deep learning algorithms, architectures, applications, and optimization methods are discussed. Various deep neural network architectures are compared based on the network model, training type, training algorithm, implementation sample, application area, sample dataset, and deep learning framework. Shrestha and Mahmood reviews in [15] the most common deep neural networks like Sparse Autoencoders, Convolution Neural Networks, Long Short-Term Memory, and Restricted Boltzmann Machines. CNN has become very well known in a variety of domains like NLP, Siamese networks, drug discovery, image recognition, and classification. This involves imposing local connectivity and extracting optimal information from raw data. Deep Learning is the most recent and popular method being used to solve many healthcare-related issues like disease diagnosis, information extraction, phenotyping, and representation learning.

TABLE 1. Comparison of proposed approach with state of the art deep learning techniques for disease prediction.

Title of the research	Method	Application Domain	Datasource	Results
Risk prediction of Acute Coronary Syndrome [22]	RSDAE-SM	Acute Coronary Syndrome (ACS)	EHR data from the Chinese PLA General Hospital	AUC= 0.868 Accuracy= 0.73
Deep Patient: Representation and prediction pf patient future using EHRs [23]	Denosing Autoencoders	Generalized disease prediction	Raw EHR dataset from Mount Sinai data warehouse	AUCROC=0.773 Accuracy=0.929
Predicting the Risk of Heart Failure with EHR Sequential Data Modeling [24]	LSTM	Congestive heart disease	Real-world dataset	AUCROC (one-hot encoding)=0.6483 AUC-ROC (vector embedding)=0.6827 ANN outperformed SVM
Comparison of ANN and SVM for prediction of chronic kidney disease [25]	Artificial Neural Network (ANN)	Chronic kidney disease	CKD dataset acquired from UCI Machine Learning Repository	ANN outperformed SVM
Big Data Predictive Analytics Model for Disease Prediction using Machine learning [26]	Naive Bayes	Heart disease	UCI Machine Learning Repository	Accuracy=0.9712
Learning to Identify Rare Disease Patients [27]	Cascade learning methodology	Rare metabolic disease lipodystrophy	Database of 2.5M anonymized EHRs arising from patient encounters with general practitioners in Europe	AUC=0.9085 Accuracy= 0.93
BEHRT: Transformer for EHRs [28]	Transformer architecture	Generalized disease prediction	Clinical Practice Research Datalink (CPRD)	AUROC=0.904
Scalable and accurate deep learning with EHRs [29]	Recurrent neural networks (long short-term memory), Attention-based TANN, Neural network with boosted time-based decision stumps	In-hospital mortality, 30-day unplanned readmission, Prolonged length of stay, Patient’s final discharge diagnoses	University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM)	AUROC In-hospital mortality =0.93 30-day unplanned readmission =0.75 Prolonged length of stay =0.85 Patient’s final discharge diagnoses =0.90 Accuracy=97.81%
Proposed Deep Similarity Learning approach	CNN_Softmax	Kidney diseases - chronic kidney disease, diabetes mellitus with kidney complications, hypertension, chronic nephritic syndrome, glomerular disorders, end-stage renal diseases	ORBDA dataset	Accuracy=97.81%

Esteva *et al.* [21] survey the application of several deep learning-based techniques that have greatly impacted the field of healthcare and medicine. The application of Computer Vision to medical imaging, Natural Language Processing to the study of EHR data, and Reinforcement Learning to robot-assisted surgery are a few of the several application areas of deep learning. A comprehensive overview of recent deep learning advancements for EHR has been presented in [1]. Deep learning approaches have been approved for achieving improved performance over traditional methods of machine learning. Various avenues of future research have been summarized in [1] which include, but are not limited to, data heterogeneity, unified patient representation, and model interpretability. Two of these challenges, data heterogeneity, and patient representation, are addressed in the current research and can be used in the development of future methods.

B. DIAGNOSIS PREDICTION

In recent years, a handful of careful research has also been done on the application of deep learning techniques for predicting diseases at an early stage and providing timely treatment. A comparative study of the state of art in the domain of deep learning and disease prediction is shown in Table 3. In [22], a corrupted version of input reconstructs to a clean repaired input using a regularized stacked denosing autoencoder. After the pre-training is complete, a softmax regression layer is added to the reconstructed feature to perform clinical risk prediction. Similarly, a fully-connected layer is added to our CNN_Softmax method after the training is complete to perform disease prediction. In other research, Miotto *et al.* [23] used stacked denosing autoencoders to automatically extract features from large EHRs and demonstrated patient representation without needing excessive human intervention. The authors showed the potential of

their method to estimate the risk of an individual contracting a particular illness. However, using an unsupervised approach like autoencoders doesn't give information about data sorting and our research involves learning similarities and grouping the patients, and finally predicting diseases. Jin *et al.* [24], identifies the relations between diagnostic events by mining time-based EHRs and then proposing a heart failure prediction neural network model using LSTM methods. One Hot encoding has been used to model the diagnostic events of patients during preprocessing and it significantly helps in improving the outcomes. Recently, big data analysis has been widely employed for disease prediction using Machine Learning. Research [25] illustrates that deep learning techniques like Artificial Neural Networks (ANNs) provide better results as compared to machine learning techniques such as Support Vector Machines (SVMs) for predicting chronic kidney disease. In [26], the proposed Big Data Predictive Analytics Model for Disease Prediction using the Naive Bayes Technique (BPA- NB) has been proven to have high accuracy in the detection of heart failure. The use of EHR data specifically, for clinical risk prediction and disease diagnosis, has seen a rise in the past few years.

Reference [27] presents a novel approach for generating accurate prediction models based on unsupervised feature selection, supervised ensemble learning, and unsupervised clustering. It uses silver standard labeled data, i.e. EHR data of patients provisionally labeled as positive for the target disease based on unconfirmed evidence. Like this study, several other studies have also employed deep learning-based models for future disease diagnosis using past EHR records. In [28], the proposed model, BEHRT, uses a transformer-based deep learning approach for prediction of future diagnosis using EHR records. BEHRT specifically uses a neural network-based model that is capable of predicting the likelihood of 301 health conditions in patients' future visits. Reference [29] has demonstrated that predictive modeling enables deep learning models to accurately predict multiple medical events using patients' EHR data. The authors achieved high accuracy for tasks such as prediction of patient mortality, 30-day unplanned readmission, prolonged length of stay, and final discharge diagnosis. Instead of standard variables, the learned model leverages several small data points unique to the individual EHR. The methods employed are computationally costly and require special expertise to be developed. The readmission dataset consists of many patient features that may have imbalanced class distributions. Therefore, an effective hospital readmission risk prediction method has been developed in [30] using the class-imbalance feature selection framework. A loss-function within a large margin method has been designed to deal with class-imbalance problems.

Several deep neural network methods have been proposed in [31]–[34] that predict the presence or absence of chronic kidney disease, perform a risk assessment of kidney disease in hypertension patients, and extracts and classify patient features. The experimental results

show that these models outperform other machine learning classifiers.

In [35], a Twitter data analytics tool named Sehaa for healthcare is implemented. It captures unstructured twitter data (Arabic language tweets) using Twitter API, then manually labels the tweets and builds different classifiers using several machine learning models such as Naive Bayes, Logistic Regression, and several other methods of feature extraction to classify the tweets and then detect different diseases. The work is an enhancement in the healthcare domain as it integrates with social media data for the betterment of public health. However, one notable limitation is that it uses manual labeling which is a very time-consuming task. Alternatively, advanced deep learning methods can be used for automatically extracting and labeling large data.

Recent advances in the Internet of Things (IoT) and big data also have brought new healthcare opportunities. The authors discuss problems linked to network latency and bandwidth in [36]. A ubiquitous health-care system is introduced, known as UbeHealth. Network traffic is predicted using big data, high-performance computing, and deep learning so that network-related problems such as latency, energy consumption, and security could be effectively resolved. In [37], a three-tier Internet of Things architecture for collecting and storing data, and developing a machine learning algorithm for early detection of heart diseases has been proposed. It uses a logistic regression-based model for predicting heart disease. The proposed model allows 5G mobile networks to transfer health data to the hospital database to take the appropriate action in emergency circumstances.

To enhance clinical decision-making and refine healthcare processes, the motivation behind such studies is to build general-purpose methods to reliably predict the duration of stay, potential disease, readmission, and mortality. Early diagnosis of health care tends to do specifically with protecting the life of people. Additionally, the detection of novel patterns will contribute to new theories and research objectives. The studied work has relied on conventional modeling techniques such as logistic regression (LR) and support vector machines (SVM), autoencoders, and long-short term memory (LSTM) using features that represent the aggregation of events or features in an observation window. In contrast, convolutional neural network (CNN) methods capture patterns and extract optimal information that is present in longitudinal data. CNN has also shown its ability to derive patient representations while building classifiers or predictors. However, traditional CNN architecture is not sufficient to fully utilize contextual information of EHRs and cannot find the relationships between different patients and provide personalized treatments.

C. SIMILARITY LEARNING

For any new patient, analyzing previous reports of patients having similar characteristics can help in identifying reference cases for predicting clinical results. In earlier research, the patient similarity was calculated by using traditional

approaches such as Cosine [38], Euclidean [39], Manhattan [40], Jaccard similarity [41], and Hamming distance [42]. These methods directly measured the similarity on raw input feature vectors which are highly dimensional, noisy, and sparse without learning the parameters on the input vector. Zhu *et al.* [43] propose supervised and unsupervised methods using a patient similarity evaluation framework to preserve the temporal properties of EHR. The result shows that supervised methods are succeeding in finding similar patients, so the current study uses the same supervised practice. In recent research [44] softmax, and triplet loss techniques, calculate similarity based on the length of the visits of a patient by appropriately representing the medical records. A softmax-based technique classifies pairwise labels into classes, and a triplet loss technique margin learns to separate the patients into positive and negative classes. The current research uses a similar softmax-based technique with an enhancement dealing with challenges such as data heterogeneity and sparsity not discussed in [44]. Wang and Sun [45] developed a weakly-supervised method of studying patient similarities that requires just a limited amount of supervisory knowledge received from the physiologists. Suo *et al.* [46] developed a patient similarity learning, named, Metric Learning with Incomplete Modalities (MeLIM) with complete as well as incomplete data. The incomplete data is said to provide more information about the characteristics and relationships between modalities. Authors used generative adversarial networks (GAN) to find relationships between modalities and a discriminator to differentiate true data and the generated data. For metric learning, they used the linear and non-linear mapping function that measures the distances between two patient samples.

The present authors, in their earlier study, evaluated the nephrology dataset through a deep learning technique, Feed Forward Network [47]. The current research mainly focuses on evaluating patient similarities which can be used for many clinical research trials, such as risk stratification (clustering patients due to their vulnerability from a medical condition), comparative efficacy analysis, and predictive modeling. This article proposes a novel deep similarity learning approach that we call “CNN_Softmax” for predictive analysis by using patient similarity on standardized EHR data. Our approach consists of two parts: patient representation learning and patient similarity learning. In patient representation learning, we use Convolutional Neural Network (CNN) to capture essential information and regular patterns in the large-scale raw EHRs and obtain a vector representation containing important features of the original patient data. In patient similarity learning, we use a Softmax-based supervised classification technique which is a Siamese-based neural network method to perform multi-class classification on learned patient representations by measuring the similarity probability between similar pairs of patients and dissimilar pairs of patients. The similarity probability between a pair of patients is measured to indicate the risk of having the same disease between two patients. After obtaining the patient

similarity information, we then perform disease prediction. The significant challenges encountered dealing with raw EHR while performing the above two tasks are heterogeneity and sparseness in data. We handle these challenges using suitable approaches, as illustrated in current research. We further compare the performance of our model with CPU and GPU systems. We acquire the data from the ORBDA dataset, which is a standardized dataset for semantic interoperability of electronic health records based on OpenEHR [19], [20]. Our experimental results show a decrease in training time for the GPU system compared to the CPU system.

IV. RESEARCH PROBLEM AND SOLUTION

EHR market analysis and forecast have become a must in the prevalent scenario. Expert decision support systems can help doctors diagnose and predict certain life-threatening illnesses. These schemes can decrease price and waiting time, and free human specialists (doctors) for further studies and reduce the mistakes and errors that people can make due to exhaustion and fatigue. This article aims to derive essential insights and knowledge from electronic health records data to predict new patient conditions.

Different deep learning architectures train the model for optimal representation of patient data. The deep learning approaches help us in finding out the essential characteristics of patients and provide personalized treatments. However, deep learning approaches are not sufficient to train a model and cannot find the relationships between different patients. Therefore, there is a need for developing methods that jointly learn the patient representations and find the relationship between the patients using pairwise similarity learning. The traditional similarity learning methods directly measure the similarity on input feature vectors without learning the parameters on the input vector.

Therefore, the current study proposes a modernized deep similarity approach to perform predictive analysis on standardized EHRs. Figure 1 shows a high-level conceptual framework of the Deep Similarity Learning approach. EHRs are first collected from the ORBDA dataset (standardized EHR dataset) [20], pre-processed for the identification and normalization of clinically significant phenotypes, and clustered into patient vectors (i.e., raw patient representation, Figure 1A). Every patient may be represented by a single vector or by a series of computed vectors, for example, predefined temporal frames. The patient vectors obtained are cleaned and pre-processed to handle sparse and heterogeneous data. We use polynomial interpolation to remove sparseness from data and datatype mapping using the OpenEHR model for the heterogeneous nature of data (Figure 1B). The vectors obtained are then used as pairwise input and fed into CNN to extract a set of high-level general descriptors using representation learning (Figure 1C).

Each layer generates a higher-level representation of the patient that is more abstract by observing patterns based on input received from a previous layer. Each patient in the dataset is then optimally represented as a one-hot patient

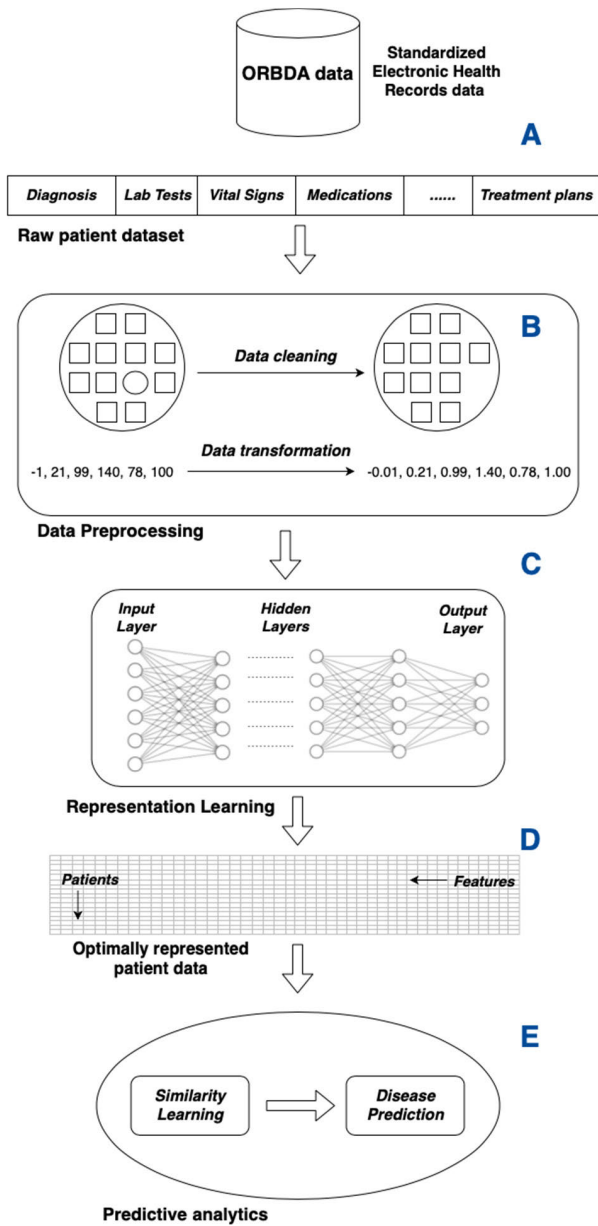


FIGURE 1. High-level conceptual framework of Deep Similarity Learning approach. (A) Raw representation of patients collected from the ORBDA dataset. (B) Pre-processing of input data. (C) Pairwise vector data modeled by supervised deep learning architecture to learn representation. (D) Optimally represented patient data. (E) Measuring similarity and performing disease prediction using pairwise vector representation.

vector using its medical features. Figure 1D shows this patient representation where each row represents a single patient and the column represents its features. The pairwise vector representation is passed to a softmax-based function to learn the similarity between patients and perform disease prediction (Figure 1E).

Section A gives the key contributions of the current research below. It also lists the methods to handle the challenges mentioned above.

A. KEY CONTRIBUTIONS

The objective of the current research is to propose a novel Deep Similarity learning approach on Electronic Health Records (EHRs) data. The similarity among patients is measured and used for diagnosing the diseases in patients. The following section illustrates the key contributions of the research.

1) HANDLING MISSING VALUES

EHR data contains sensitive patient information with missing values. We apply interpolation to remove sparseness. Interpolation is a more robust technique to estimate values between two known points, offering continuity between segments.

2) HETEROGENEITY

Healthcare data is being collected continually in a wide variety of data types, e.g., clinical notes, lab tests, medical images, demographic profiles, drug administrations, diagnoses. The OpenEHR model used in the current research can accommodate these wide varieties of data types.

3) OPTIMAL REPRESENTATION LEARNING

A deep learning architecture, namely, CNN, is used for representing longitudinal EHR data and obtaining a useful representation of input features. CNN maps input features to output data (1-D data) without losing essential characteristics.

4) SIMILARITY LEARNING

A softmax-based supervised classification technique classifies pairwise labels into classes and obtains a similarity probability among the pair of patients indicating the possibility of developing the same disease. Then the training patients are ranked according to their similarity score from new patients in ascending order.

5) ONE-HOT ENCODING

The one-hot encoder performs category “binarization” and includes it as a feature or label to train the model. The one-hot encoding process is used for better performance, by which categorical labels convert into 1s and 0s where 0 indicates non-existence and 1 indicates the label’s existence.

6) SEMANTIC INTEROPERABLE DATASET

We are working on the openEHR compliant dataset, which is a standard for semantic interoperability. OpenEHR is a standard that follows a dual model approach (consisting of a reference model and an archetype model). The main objective of the standardized data is to provide interoperability of EHRs among medical institutions. Various organizations such as HL7, ISO, and openEHR are working on standards for semantically interoperable EHRs [20]. However, the current study’s effort to perform disease prediction on a dataset from one of the semantically interoperable standards (openEHR) is the first attempt.

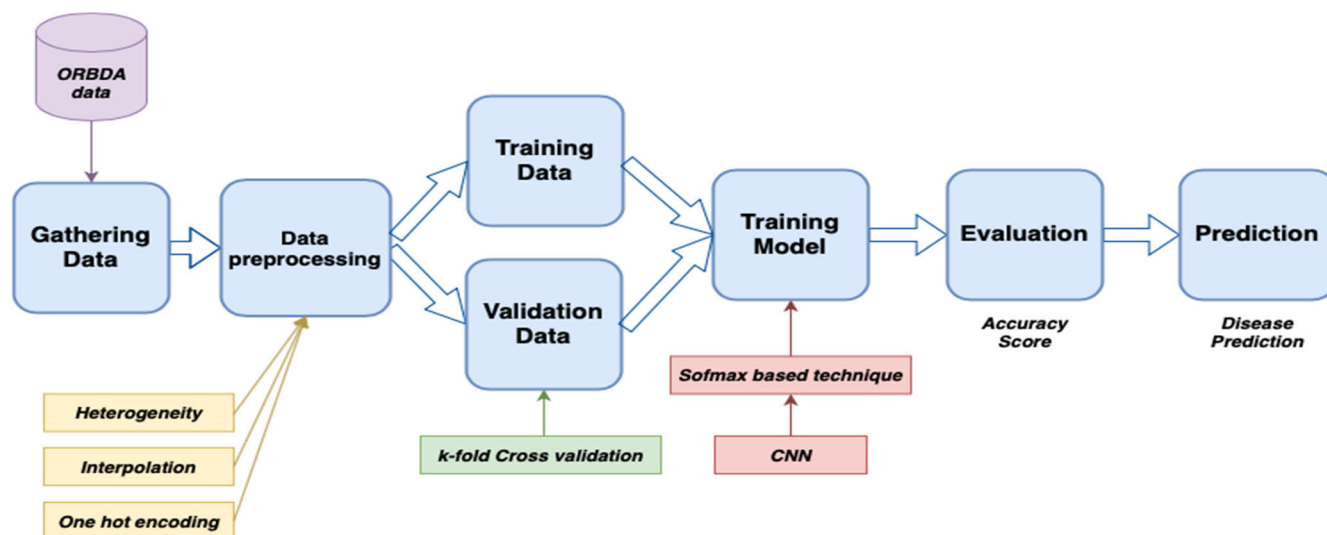


FIGURE 2. The architecture of the modernized Deep Similarity Learning approach.

V. METHOD

This section describes in detail the methodology of our proposed modernized Deep Similarity Learning approach to electronic health record data. Figure 2 presents the architecture of the Deep Similarity Learning approach.

The various steps performed on our model are data gathering, data preprocessing, splitting the data into training and validation data, training the model, evaluation of the model, and disease prediction. We start by gathering the nephrology data from the ORBDA dataset. The data is then prepared by removing missing values, maintaining heterogeneity in the data, and using one-hot encoding for the output values. We train our model with the developed novel CNN_Softmax methodology, where we use CNN for optimal representation of longitudinal data and a softmax-based supervised classification technique to measure similarity among patients.

Further subsections are the elaboration of the architecture diagram. Section A is a data gathering and data preprocessing step. Under Section A, sub-section 1 gives the underlying representation of data collected, and sub-section 2 discusses handling the heterogeneous nature of data. Sub-section 3 provides details about handling missing values. We are using interpolation techniques for this purpose. Section B describes how training is performed on the model using a Deep Similarity approach. For training our model and calculating similarity, we use CNN and softmax based technique, as explained in sub-section 1 and 2 under section B, respectively. We finally perform disease prediction after training the model to diagnose the disease from which a patient might be suffering, as described in subsection 3.

A. DATA GATHERING AND PREPROCESSING

We start by gathering the nephrology data from the openEHR benchmark dataset, which is a publicly available dataset [19] for assessing the performance of EHRs. The data is then

prepared by removing missing values, maintaining heterogeneity in the data, and using one-hot encoding for the output values.

1) REPRESENTATION

We assume there are N patients, then the nth patient p_n having kidney disease can be represented by multiple input features as x₁, x₂, . . . , x₁₂ described in Table 2. Medical codes recorded as the International statistical classification of diseases and related health problems (ICD-10) captures the disease diagnosed or treatment received by the patient [48].

TABLE 2. Input features in the dataset.

Representation	Input features
x ₁	HIC_antibodies
x ₂	HIV
x ₃	HbsAg
x ₄	Age
x ₅	healthcare_unit
x ₆	owner_id
x ₇	Procedure
x ₈	reason_for_discharge
x ₉	State
x ₁₀	urea_reduction_rate
x ₁₁	venous_fistula_amount
x ₁₂	Volume

c₁, c₂, . . . , c_{|C|} ∈ C denotes these medical codes, where |C| denotes the number of different codes related to kidney disease. main_diagnosis denotes all the codes corresponding to the output label or prediction in the EHR data. There are ten output classes shown in Table 3.

Encoding Categorical labels: For efficient implementation of our algorithm, we convert our categorical labels into a one-hot encoded matrix. In one-hot encoding, a column with categorical data splits into multiple columns, and each value

TABLE 3. Output categorical values.

Class	ICD10 codes	Disease	No. of records
0	E10_2	Type 1 diabetes mellitus with kidney complications	972
1	E14_2	Unspecified diabetes mellitus with renal complications	1814
2	I10	Essential primary hypertension	3507
3	I12_0	Hypertensive chronic kidney disease	11112
4	N03_9	Chronic nephritic syndrome with unspecified morphologic changes	3623
5	N08_3	Glomerular disorders in diseases classified elsewhere	3047
6	N08_8	Glomerular disorders in diseases classified elsewhere	1218
7	N18_0	Chronic kidney disease (CKD)	2146964
8	N18_8	End-stage renal disease	2323
9	N18_9	Unspecified Chronic kidney disease	20473

TABLE 4. Example of One-hot encoding of categorical labels.

	I10_2	N03_9	N08_8	N18_8
1				
2	1,	0,	0,	0
3	0,	1,	0,	0
4	0,	0,	1,	0
5	0,	0,	0,	1

is replaced by 1s and 0s, which has been label encoded depending on which column has which value. Table 4 shows an example of a one-hot encoding of categorical labels. For example, class I10_2 is one of the categorical labels and encoded in one-hot encoding as [1,0,0,0].

2) HETEROGENEOUS NATURE OF EHR DATA

Healthcare data is being collected continually in medical informatics in a wide variety of data types, e.g., clinical notes, lab tests, medical images, demographic profiles, drug administrations, and diagnoses. The OpenEHR model can express these large varieties of data types (greater than 20) into archetype-based data types, namely DV_QUANTITY, DV_BOOLEAN, DV_CODED_TEXT, DV_COUNT, DV_DATE, DV_DATE_TIME, DV_TEXT. [49]. We further represent these archetype-based data types into a more straightforward representation of common data types such as string, boolean, number (float or integer), and date as given in Table 5.

3) HANDLING MISSING DATA VALUES

Missing data can be handled by conventional approaches like imputing missing values using mean or median or deleting listwise, pairwise records with missing values. Since we are dealing with medical data, it is not preferred to handle missing numerical values by replacing them with mean or any other substitution methods [46]. These may lead to a faulty diagnosis, which may be fatal to a patient and also introduce biased estimates. There exist more robust approaches to handle EHR missing data [50], such as interpolation [51],

TABLE 5. Handling and mapping heterogeneous data.

Datatypes in Standard data (Archetype based)	Mapped Data Types
DV_QUANTITY	Float
DV_BOOLEAN	Boolean
DV_CODED_TEXT	String
DV_COUNT	Integer
DV_DATE	String
DV_DATE_TIME	String
DV_PROPORTION	Float
DV_TEXT	String

expectation-maximization [52], multiple imputations [53], and maximum likelihood [54]. Table 6 gives the advantages and limitations of the methods to handle missing data.

We used interpolation to process EHR missing data. Interpolation is a mathematical method to estimate missing values between two known points. The simplest type of interpolation is a linear interpolation that makes a mean between the values before and the value after the missing data value. Here, we use polynomial interpolation because linear interpolation results in discontinuities at each point, and polynomial interpolation offers continuity between segments. Also, any function continuous over a closed interval can be well approximated arbitrarily by polynomials.

B. DEEP SIMILARITY LEARNING APPROACH

This section elaborates on performing predictive analysis using a novel Deep Similarity learning approach, CNN_Softmax. CNN extracts important patterns within input

data, and a softmax-based technique measures the similarity between patients to perform disease prediction.

1) CONVOLUTIONAL NEURAL NETWORKS

A matrix containing patient data as input features appears as a matrix containing pixel values of an image. We apply a one-sided convolution operation in place of using a two-dimensional convolution as there exists no relationship between output categorical data [44]. We get the number of filters or kernels within the convolutional layer by $m = ab$, where a is the size of different filters, and b is the number of filters per size. A filter matrix filter is given by, $W_e \in R^{k \times d}$, where k is the size and d denotes the dimension of the input vector. A concatenation of filter and feature vector from x_i to x_{i-k+1} is represented by a value e_i and is given by the equation,

$$e_i = \text{ReLU}(W_e \cdot x_{i:i+k-1} + b_e) \quad (1)$$

where ‘ \cdot ’ is the dot product and $b_e \in R$ is a bias term, and ReLU is an activation function named Rectified Linear. This filter is moved with a stride equal to 1 to each possible window of a feature vector to obtain a feature map $e = \{e_1, e_2, \dots, e_{t-k+1}\}$, where $e \in R_{t-k+1}$. Due to m total filters present, we get m feature maps. It is possible to add more layers to the input layer for making the convolutional step more efficient. Each layer can be linked to different filters. Hence, we can extract different features from the original patient matrix [55].

The output of the convolutional layer is forwarded to the pooling layer over a feature map as $\hat{e} = \max\{e\}$, where \hat{e} denotes the maximum value of a specific filter. There exist two kinds of pooling, max pooling, and average pooling. Max pooling gives the maximum value from the input covered by the filter, whereas average pooling gives the average of all the values from the input covered by the filter. We use max-pooling here to store the essential information for every feature map. For each region represented by the filter, the max of that region is taken and created a new output matrix where each element is the maximum of the original matrix. Max-pooling discards noise activations performing dimensionality-reduction. Concatenating pooled output from all the filters produces a vector representation. Adding fully connected layers produces combinations of the different high-level features from the output of the convolutional and pooling layer. The learned vector $k \in R^m$ is the vector representation of the initial matrix consisting of patient input medical features.

Activation Function: The activation functions calculate the sum of input with a concatenation of weights and biases, which decides the activation of a neuron. The output of a neuron f can be shown as a function of its input x is with $f(x) = \tanh(x)$ or $f(x) = (1 + e^{-x})^{-1}$ called tanh and sigmoid functions, respectively. The range of a sigmoid function is between 0 and 1, whereas that of tanh function is between -1 and $+1$. These saturating non-linear activation functions squash the input giving slower training results.

The Rectified Linear Unit activation function (ReLU) is a faster learning non-saturating activation function that gives better performance in deep learning than tanh and sigmoid activation functions [56]. The equation, f , provides the ReLU activation function as

$$f(x) = \max(0, x) \quad (2)$$

where the range of the function lies between 0 and x .

The values that are less than zero are replaced by the ReLU function to repair the input values and remove the decreasing gradient problem seen in other activation functions.

2) SIMILARITY LEARNING

After successfully representing the EHRs as vectors in the representation phase, the next step is to calculate the similarity among different representations of EHRs. The use of the traditional similarity learning methods (cosine, Euclidean) directly measures the similarity on input feature vectors without learning the input vector parameters. These traditional methods do not perform well on original data, which is highly dimensional, noisy, and sparse. Therefore, we follow an approach used in [57] to measure the similarity between patients by first learning the parameters on the input value.

Siamese-based classification [58] using the softmax calculates the similarity between any two pairs of objects. A bilinear distance given by the following equation calculates the similarity,

$$S = k_i M k_j \quad (3)$$

here k_i and k_j corresponds to the vector representation of patient i and patient j that we received as an output from convolutional layers, and $M \in R^{m \times m}$ is the symmetric matching matrix optimized during training.

The matching matrix is written as $M = L^T L$ to ensure similarity of M . We also convert k_i and k_j into a single vector given by

$$K = W_k k_i \oplus W_k k_j \quad (4)$$

where $W_k \in R^{m \times m}$, and \oplus denotes a bitwise addition. An output probability \hat{y} is obtained by concatenating K and S and feeding it into a fully connected softmax layer. The value of \hat{y} is a float value between 0 and 1. \hat{y} has a value 1 if patients i and patient j have the risk of getting the same disease; otherwise, \hat{y} has a value 0. During the training, all the model parameters are simultaneously updated. Figure 3 shows the overall architecture for optimal representation using deep learning technique, CNN, and then measuring pairwise similarity using a softmax-based technique. The patient representations are converted using CNN. Softmax-based technique is added later to calculate the similarity vector. Algorithm 1 describes the CNN_Softmax methodology in detail.

3) DISEASE PREDICTION

After training the model, we perform disease prediction to predict the disease from which a patient might be suffering.

Algorithm 1 CNN_Softmax**Input data:** $x_i \forall i \in 1, 2, 3, \dots, n$ // x_i be a feature input vector of n patients

1. Apply a one-sided convolution operation involving a filter matrix $W_e \in \mathbb{R}^{k \times d}$ and a bias term b_e .

$$e_i = \text{ReLU}(W_e \cdot x_{i:i+k-1} + b_e)$$

// k be the size of the input vector, d be its dimension, ReLU is an activation function

2. Apply max pooling to the output of the convolution layer.

$$\hat{e} = \max\{e\}$$

// \hat{e} denotes the maximum value of each feature map obtained in step 1.

3. Concatenate output from pooling layer to form vector representation, $k \in \mathbb{R}^m$

// m is the number of filters

4. Calculate similarity score using,

$$S = k_i M k_j$$

// k_i and k_j are the vector representation of patient i and patient j . M is the symmetric matrix

5. Convert k_i and k_j to a single vector,

$$K = W_k k_i \oplus W_k k_j$$

// $W_k \in \mathbb{R}^{m \times m}$

6. Concatenate K and S and feed into a fully connected softmax layer to obtain $\hat{y} \in [0,1]$

// \hat{y} is the output similarity probability

We calculate the distances between each training patient and each new patient for classification purposes after mapping the representation matrix to another hyperplane. Then, according to their similarity score, we rank training patients from new patients in ascending order.

VI. EXPERIMENTAL RESULTS

In this section, we give the details of experimental results by evaluating our model on standardized EHR data. This study is the first attempt to work on standardized (openEHR) data, to the best of the author's knowledge.

A. DATASET DESCRIPTION

The dataset used in this study is known as ORBDA (OpenEHRBenchmark Dataset). OpenEHR is an open standard that defines specifications in health informatics for storage, retrieval, and sharing of Electronic Health Records (EHRs) [19]. The openEHR architecture aims to allow future-proof interoperable Electronic Health Records to be applied using accessible open designs and content

descriptions using a common technical reference system and an archetype model at the level of clinical interpretation, bringing semantic and syntactic interoperability to the EHR setting.

The dataset of ORBDA consists of hospitalization data and high complexity procedures data. The data is encoded as two kinds of records - AIH having 5.7 million records and APAC having 9.5 million records. An entry in the AIH database is created whenever a medical institution generates a hospitalization request. While on the other side, medical service providers create documents in the APAC database to record approved high complexity processes for accounting purposes. While AIH data is recorded in a single file, occurrences recorded in the APAC database are further split into six distinct classifications –bariatric surgery, chemotherapy, medicine, nephrology, radiotherapy, and outpatient miscellaneous. The dataset is filtered into a nephrology dataset (containing 5.07% records) from the APAC database and used in evaluating the model for the current study. Figure 4 provides an outline of ORBDA separated into AIH and APAC. The portion of ORBDA highlighted (Nephrology) is used for analysis and disease prediction.

The dataset chosen for this research contains 2.2 million records. We consider twelve input features (Table 2) for this study. We are excluding other features because of fewer entries available. Using the main_diagnosis label, which contains ten output classes following ICD10 codes (Table 3), we make the output or the prediction.

B. SOFTWARE AND HARDWARE CONFIGURATION

Considering the evolving nature of EHRs data, we perform the experiments on CPU as well as GPU. For performance evaluation and metric calculation on CPU, we use the following system- OS Ubuntu 18.04LTS x64, Intel Core i5-4210U CPU @1.70GHz (4 cores), 4GB DDR3 RAM. We also perform our experiments on another system with OS Ubuntu 19.04 LTS x64, Intel Xeon CPU E5-2620 @2.1GHz (6 cores), 16GB DDR3 RAM. For performance evaluation and metric calculation on GPU, we use NVIDIA QuadroK2200 @1045MHz (4GB GDDR5 VRAM). We implement the proposed approach with TensorFlow [59] and Keras [60]. Parameters of the model are optimized using Adam [61].

C. PERFORMANCE EVALUATION

This section describes the results obtained after evaluating the model on a real EHR dataset. The dataset used belongs to nephrology, and we partition it into training and test data in a ratio of 0.80:0.20 randomly. To assess the performance of CNN_Softmax on the nephrology dataset for predicting disease, we measure Accuracy, Recall, Precision, and F1 score, as these are the standard measures that help to determine how good a learning model is. We evaluate our model using macro-averaging [62] since we are performing multi-class classification. The following are the formulas to calculate these measures. (5)–(8), as shown at the bottom of the page.

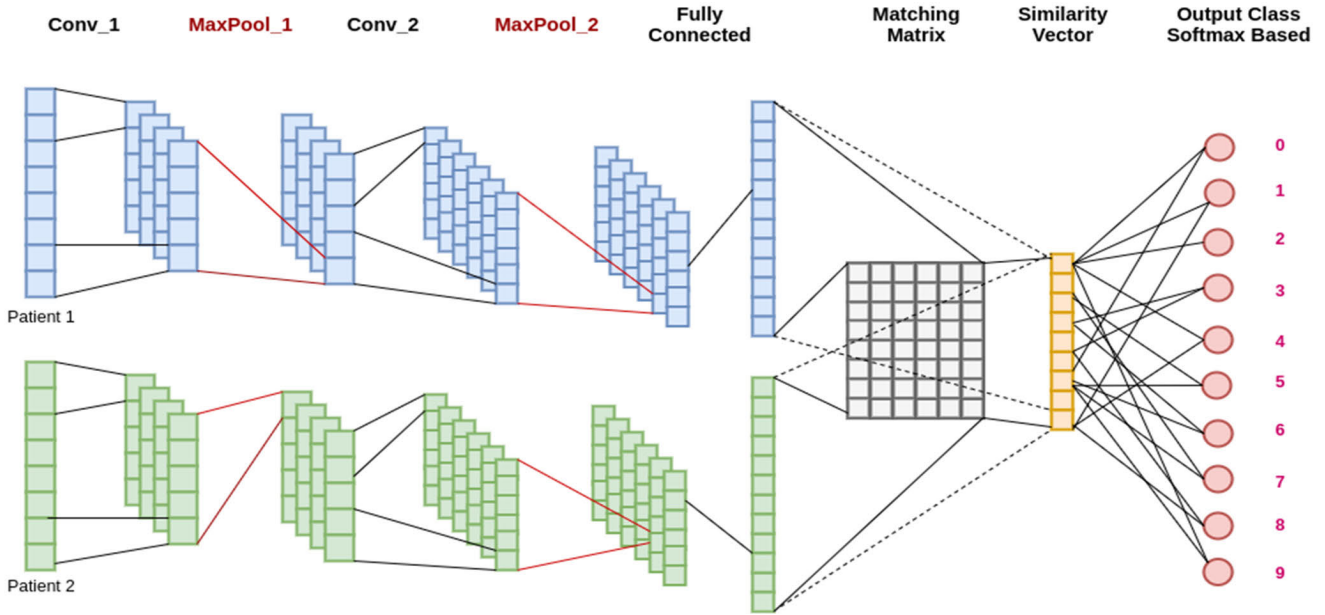


FIGURE 3. Detailed architecture of the CNN_Softmax method for learning similarity between a pair of patients.

We performed experiments in two parts. In the first part, we implemented conventional supervised machine learning and deep learning techniques for the classification of diseases to analyze the performance measure. In the second part, we concatenated a deep learning technique that optimally represents the patient data with different similarity learning methods to analyze the performance measure of different deep similarity learning methods.

Table 7 shows the comparative performance measure results of different supervised machine learning techniques such as Logistic Regression, Naïve Bayes, KNN, and Decision trees. The accuracy of all conventional machine learning techniques is above 94% which validates the effectiveness of applying these techniques for disease prediction. However, it is important to note that these accuracy values cannot be used as a criterion to make precise decisions about a complicated problem like patient similarity learning which involves finding optimal representations, often at multiple levels, where higher-level features are learned in terms of lower-level features [2], [63] and then using these representations to learn pairwise similarity.

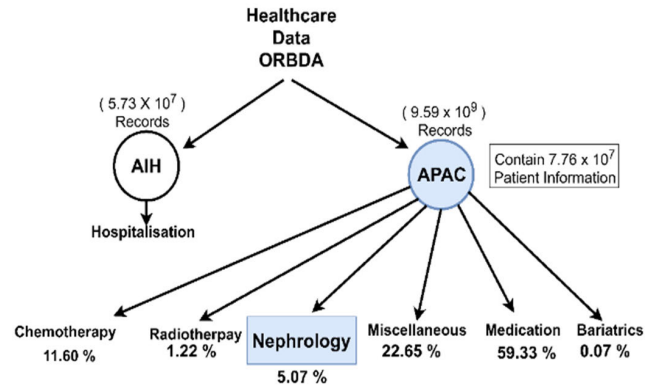


FIGURE 4. ORBDA overview with filtered problem domains.

To validate the results of optimally representing the health data for each patient, we trained the model using three deep learning architectures, ANN (MLP), CNN, and RNN (LSTM). As can be seen from Table 8, the MLP model achieved 98.74% training accuracy, whereas CNN and RNN achieved 98.72% and 96.71% training accuracy, respectively.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True\ Negative} \tag{5}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{6}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{7}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

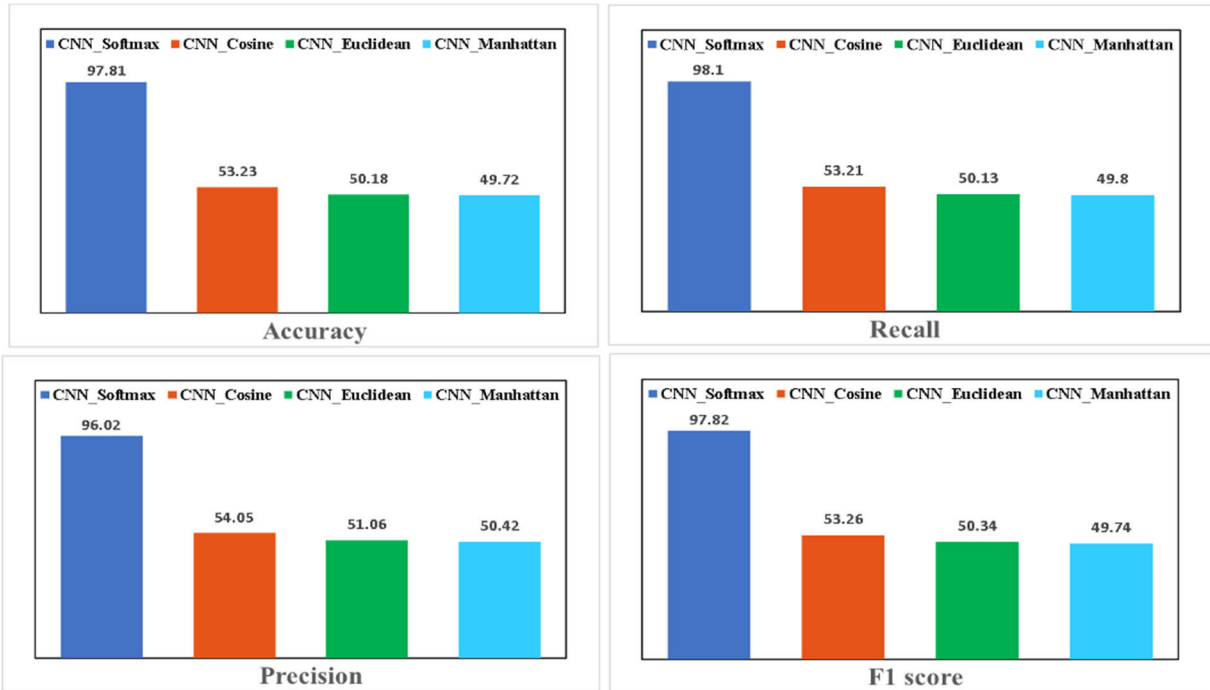


FIGURE 5. Performance of different similarity learning methods based on multiclass Accuracy, Recall, Precision, and F1 score. CNN_Softmax outperformed the traditional similarity learning methods.

Although MLP gives slightly better results than CNN, the layers in CNN are sparsely connected rather than fully connected (as in MLP), hence making the CNN architecture go deeper rather than growing bigger. Also, CNN is considered best for a classification-based supervised task where medical data features need to be optimally represented by finding the patterns using filters. Exploring these deep learning architectures helped us in finding out the essential characteristics of patients. However, deep learning approaches are not sufficient to train a model and cannot find the relationships between different patients [64]. Therefore, there is a need for developing methods that jointly learn the patient representations and find the relationship between the patients using pairwise similarity learning.

To optimally represent the health data for each patient, we use 1D CNN, which effectively derives important features from fixed-length segments of the complete dataset. For similarity learning, we train the model to learn a similarity degree among patients using a softmax-based technique. To validate the performance of our novel modernized Deep Similarity learning approach, CNN_Softmax, we compare it with traditional similarity metric learning methods. We first represent the patient input data using CNN and then use the conventional similarity learning methods Manhattan, Euclidean, and Cosine, to learn the similarity among pairs of patients.

Table 9 shows that CNN_Manhattan, CNN_Euclidean, CNN_Cosine do not perform well because they directly measure the similarity on raw input feature vectors which are highly dimensional, noisy, and sparse, without learning the parameters on the input vector. The performance of different

similarity learning methods based on multiclass Accuracy, Recall, Precision, and F1 score is shown in Figure 5. Our novel approach CNN_Softmax outperforms state-of-the-art methods.

Figure 6 illustrates the accuracy of CNN_Softmax on data it is constructed on and the accuracy of the model on data it hasn't seen. The training accuracy and testing accuracy achieved is 97.81% and 97.82%, respectively which shows how well the model has generalized to new unseen data and it did not overfit. Similarly, Figure 7 shows the training and test loss indicating how bad the model's prediction is on a single example. Our model utilizes cross-entropy loss and the result demonstrates that pairs are correctly classified. The results illustrate that CNN_Softmax performs disease prediction with effective accuracy and reduced loss.

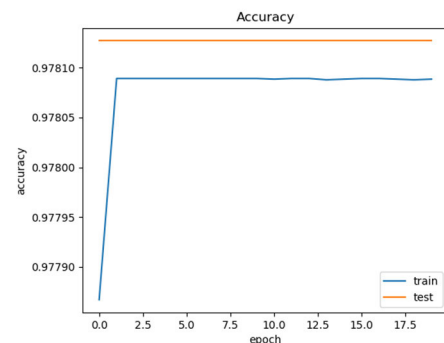


FIGURE 6. Accuracy.

TABLE 6. Methods to handle missing data.

Method	Advantages	Limitations
Listwise deletion, mean filling	Easy to implement; fast	Adds no new information; reduces variance; introduce biases
Nearest neighbor, Hot deck	Easy to interpret; consistent to cross user	Reduces variance; introduce biases
Interpolation (linear, quadratic, spline, cubic)	More accurate; direct estimation based on neighbors	Does not give relationships between different features
Model-based filling (expectation-maximization, multiple imputations, maximum likelihood)	Improves precision; resistant to outliers; Results in robust statistics	Does not work for mechanisms like MCAR, MAR, and MNAR

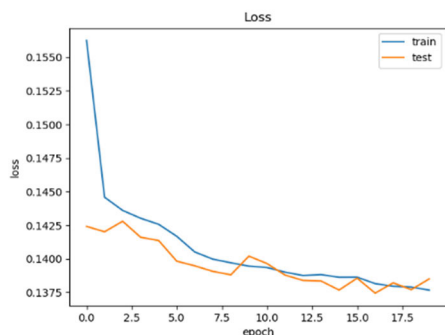


FIGURE 7. Loss.

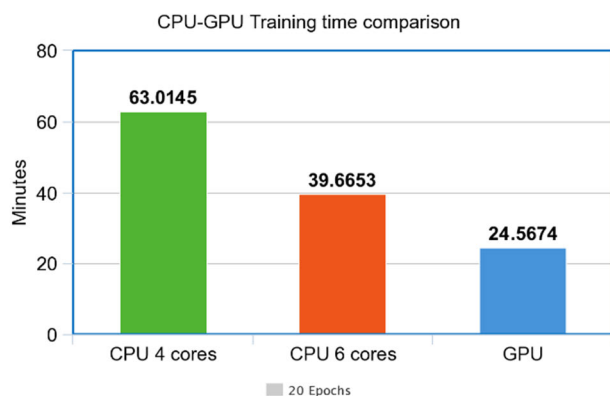


FIGURE 8. CPU-GPU Training time.

The current study also compares the training time for performance evaluation between four-core CPU, six-core CPU, and GPU for 20 Epochs in addition to calculating the accuracies. Figure 8 gives a comparison between CPU and GPU training time. Y-axis denotes the time taken to train the model in minutes for CPU and GPU systems as compared to training on CPU systems. We can see that there is a decrease in the training time for GPU.

To validate the performance of using polynomial interpolation for handling missing data in EHR, we compare it by measuring the performance of our model based on different methods of dealing with missing values. Table 10 shows

TABLE 7. The performance measure of conventional machine learning techniques.

Method	Accuracy	Recall	Precision	F1 score
Logistic Regression	0.9683	0.9635	0.9472	0.9576
Naive Bayes	0.9465	0.9454	0.9371	0.9489
KNN	0.9709	0.9706	0.9683	0.9693
Decision trees	0.9742	0.9718	0.9703	0.9722

TABLE 8. The performance measure of conventional deep learning architectures.

Deep Learning architecture	Accuracy	Loss
RNN	0.9671	0.1528
CNN	0.9872	0.1270
MLP	0.9874	0.0800

TABLE 9. The performance measure of deep similarity learning methods.

Method	Accuracy	Recall	Precision	F1 score	Loss
CNN_Manhattan	0.4972	0.4980	0.5042	0.4974	0.0153
CNN_Euclidean	0.5018	0.5013	0.5106	0.5034	0.0797
CNN_Cosine	0.5323	0.5321	0.5405	0.5326	0.0859
CNN_Softmax	0.9781	0.9810	0.9602	0.9782	0.1386

the accuracy score, recall, precision, and F1 score using macro-averaging obtained by taking the missing values by dropping null values, and applying interpolation on our dataset.

Though we get slightly higher results by dropping the null values, it is not the right method as many other essential values associated with it in a row also get lost. We also perform k-fold cross-validation on our dataset to evaluate our model. We choose the value of k as ten because this value generally results in a low bias and a modest variance [65]. Using k-fold cross-validation, our model achieved an accuracy of 97.80%. Hence, CNN_Softmax using interpolation performs disease prediction by obtaining vital information about patients and calculating a similarity score.

TABLE 10. The measure of performance of the model based on methods dealing with missing values.

CNN_Softmax	Accuracy	Recall	Precision	F1 score	Loss
With missing values	0.7616	0.77	0.75	0.76	0.0385
Dropping missing values	0.9874	0.99	0.98	0.98	0.0918
Interpolation on missing values	0.9781	0.98	0.96	0.97	0.1386

TABLE 11. Comparison of proposed approach with state of the art deep learning techniques for kidney disease prediction.

Title of the research	Deep Learning technique	Application Domain	Datasource	Results
Prediction of Chronic Kidney Disease [31]	Artificial Neural Network	Chronic Kidney Disease	CKD dataset acquired from UCI Machine Learning Repository	Accuracy=97.76%
Kidney disease prediction in hypertension patients using EHR [32]	Initial modeling of the problem as a binary classification task, then proposal of a hybrid neural network incorporating BiLSTM and autoencoder networks	Kidney Disease	Raw EHR dataset from hospitals in China.	The proposed model received 89.7 % accuracy
Analysis of AI approaches to improve kidney care [33]		Acute Kidney Injury and Chronic Kidney Disease		AI has great potential for improving kidney care
Comparison of ANN and SVM for prediction of chronic kidney disease [25]	Artificial Neural Network (ANN)	Chronic Kidney Disease	CKD dataset acquired from UCI Machine Learning Repository	ANN outperformed SVM
Analysis and prediction of Chronic Kidney Disease using simple hardware sensing module and Deep Learning architecture [34]	Convolutional Neural Network-Support Vector Machine (CNN-SVM) hybrid deep learning approach	Chronic Kidney Disease		Accuracy=96.59%
Prediction of Kidney diseases [47]	Feed Forward Neural Network	Kidney Disease	ORDBA dataset	Accuracy= 98.7%
Proposed Deep Similarity Learning approach	CNN_Softmax	Kidney Diseases - chronic kidney disease, diabetes mellitus with kidney complications, hypertension, chronic nephritic syndrome, glomerular disorders, end-stage renal diseases.	ORBDA dataset	Accuracy=97.81%

VII. DISCUSSION

This section discusses and compares our proposed deep similarity approach with the application of deep learning methods for the prediction of kidney-related diseases in recent years. Table 11 presents a clear comparison of the different state-of-the-art deep learning models that have worked on analyzing and predicting kidney diseases with their results. After rigorous analysis, we observe that the existing approaches developed by the researchers mainly work on analyzing and predicting chronic kidney diseases.

The current study proposes a Deep Similarity learning approach that predicts different kidney-related disorders, including chronic kidney disease, diabetes mellitus with kidney complications, hypertension, chronic nephritic syndrome, glomerular disorders, and end-stage renal diseases. Also, many deep learning approaches have been utilized for disease prediction in the healthcare domain to find out the

essential characteristics of patients and provide personalized treatments. The present authors in their earlier work [47] used Feedforward Neural Networks on the ORBDA dataset as a preliminary step towards performing prediction of kidney diseases. Although accuracy is high in [47], authors do not study any traditional similarity learning methods or propose any novel method to evaluate patient similarities. Also, various challenges associated while handling large raw EHRs are not discussed in [47]. The current research is an extension of the work and the methodology adopted in reference [47]. It is important to note that the main focus of current research is to evaluate patient similarities which can be used for many clinical research trials, such as risk stratification (clustering patients due to their vulnerability from a medical condition), comparative efficacy analysis, and predictive modeling. Convolutional Neural Networks, considered as a type of feed-forward neural network with sparsely connected layers, are used to optimally represent patient data which is then fed

to a softmax function which learns the similarity between pairs of patients. This can be later used to compare or match new samples from previously-unseen categories. Measuring similarity could contribute to building more accurate or efficient classifiers which will help physicians identify similar cases, and the probability that the patient may be effectively treated will be significantly increased. Therefore, there is a need for developing methods that jointly learn the patient representations and find the relationship between the patients using pairwise similarity. The current study proposes such a novel approach, Deep Similarity learning, which uses a methodology, CNN_Softmax.

VIII. CONCLUSION

Patient's risk assessment of developing a particular disease and providing personalized healthcare is an important research area that helps to provide better diagnosis and treatment to patients. An initial work, using three deep learning architectures, ANN (MLP), CNN, and RNN (LSTM), trains the model for optimal representation of patient data. The idea is first to derive important local patient information using representation learning and then measure similarity among patients using a similarity learning method. The objective of learning the similarity between patients is to discover how similar any pair of patients is, based on their medical information. Healthcare datasets propose several similarity learning methods. We use a softmax-based supervised classification technique for obtaining similarity probability between the patients. Significant challenges encountered during measuring similarity for large EHRs are heterogeneity, representation, and sparsity. The OpenEHR model is used, which can express different data types and thus handles the heterogeneous nature of the dataset. The current study uses CNN for the representation of longitudinal data and to obtain an effective representation that contains essential characteristics of the original patient data. To handle the sparseness of data, we use the polynomial interpolation method. Hence a Deep Similarity learning-based approach, i.e., CNN_Softmax, is proposed to perform predictive analysis of kidney-related diseases based on patient medical features. A standardized nephrology dataset is used to train the model, and it achieves an effective accuracy of 97.81%, which outperforms the traditional similarity learning methods. Our experimental results also show a decrease in training time for the GPU systems compared to CPU systems. This research may help in an epidemiological scenario such as COVID-19.

We also here mention some limitations of our research which illustrates some opportunities for enhancement in future methods. Due to irregular visits of patients and insufficient recordings, we used high complexity procedure data to capture the medical features of patients in a single vector representation and did not include hospitalization data, which gives temporally sequenced patient data in a time interval. The proposed approach can also be applied to the medical image dataset to study outcome prediction. The Deep Similarity learning model can further be enhanced where large

complex models and training on large datasets becomes necessary for improving the performance. A database may be used for Distributed Machine learning. DBMS-based learning allows for trivial scaling to large data sets and particularly large models where different computing units work on different sections of a model which may be too complex to fit in with RAM. Other applications of Deep Similarity learning in the field of healthcare, like disease inference, personalized medicines, trajectory analysis, drug discovery, and clinical trial patient recruitment can be advanced and expert knowledge can be incorporated. The different scenarios of the dataset shift can be investigated in a detailed manner.

REFERENCES

- [1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.
- [2] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 195, Dec. 2019.
- [3] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, Jun. 2012.
- [4] Q. Wu, A. Boueiz, A. Bozkurt, and A. Masoomi, "Deep learning methods for predicting disease status using genomic data," *J. Biometrics Biostatist.*, vol. 9, no. 5, 2018.
- [5] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [6] N. Polyzotis, S. Roy, and S. Whang, "Data lifecycle challenges in production machine learning: A survey," *ACM SIGMOD Rec.*, vol. 47, no. 2, pp. 17–28, 2018.
- [7] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 1, pp. 198–208, Jan. 2017.
- [8] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J. Biomed. Informat.*, vol. 46, no. 5, pp. 830–836, Oct. 2013.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [10] T. Ruan, L. Lei, and Y. Zhou, "Representation learning for clinical time series prediction tasks in electronic health records," *BMC Med. Inform. Decis. Making*, vol. 19, no. 8, p. 259, 2019, doi: [10.1186/s12911-019-0985-7](https://doi.org/10.1186/s12911-019-0985-7).
- [11] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 539–546, doi: [10.1109/CVPR.2005.202](https://doi.org/10.1109/CVPR.2005.202).
- [12] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: leveraging patient similarity and drug similarity analytics," *AMIA Summits Transl. Sci. Proc.*, vol. 2014, pp. 132–136, Apr. 2014.
- [13] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Informat.*, vol. 69, pp. 218–229, May 2017.
- [14] R. Miotto and C. Weng, "Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials," *J. Amer. Med. Inform. Assoc.*, vol. 22, no. e1, pp. e141–e150, Apr. 2015.
- [15] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.
- [16] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018.
- [17] S. Pouyanfar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, p. 92, 2018.
- [18] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

- [19] (2019). *Generation of a Public Base for Evaluation of Persistence Mechanisms of Electronic Health Records Systems Based on the openEHR Foundation Specifications*. L@MPADA/UERJ-InformáticaMédica. Lampada.uerj.br Accessed: Apr. 17, 2019. [Online]. Available: <http://www.lampada.uerj.br/en/orbda/>
- [20] S. Sachdeva and S. Bhalla, "Semantic interoperability in standardized electronic health record databases," *J. Data Inf. Qual.*, vol. 3, no. 1, pp. 1–37, Apr. 2012.
- [21] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.
- [22] Z. Huang, W. Dong, H. Duan, and J. Liu, "A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 956–968, May 2018.
- [23] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, no. 1, p. 26094, May 2016.
- [24] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, "Predicting the risk of heart failure with EHR sequential data modeling," *IEEE Access*, vol. 6, pp. 9256–9261, 2018.
- [25] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101–111, Jun. 2019.
- [26] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of big data predictive analytics model for disease prediction using machine learning technique," *J. Med. Syst.*, vol. 43, no. 8, p. 272, Aug. 2019.
- [27] R. Colbaugh, K. Glass, C. Rudolf, and M. Tremblay, "Learning to identify rare disease patients from electronic health records," in *Proc. AMIA Annu. Symp.*, Dec. 2018, pp. 340–347.
- [28] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, p. 7155, Dec. 2020.
- [29] A. Rajkomar, E. Oren, and J. Dean, "Scalable and accurate deep learning with electronic health records," *npj Digit. Med.*, vol. 1, no. 1, p. 18, 2018.
- [30] G. Du, J. Zhang, Z. Luo, F. Ma, L. Ma, and S. Li, "Joint imbalanced classification and feature selection for hospital readmissions," *Knowl.-Based Syst.*, vol. 200, Jul. 2020, Art. no. 106020.
- [31] H. Kriplani, Himanshu, B. Patel and S. Roy, "Prediction of chronic kidney diseases using deep artificial neural network technique," in *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, vol. 31. Cham, Switzerland: Springer, 2019, pp. 179–187.
- [32] Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records," *BMC Med. Informat. Decis. Making*, vol. 19, no. S2, p. 51, Apr. 2019.
- [33] P. Rashidi and A. Bihorac, "Artificial intelligence approaches to improve kidney care," *Nature Rev. Nephrol.*, vol. 16, no. 2, pp. 71–72, Feb. 2020.
- [34] N. Bhaskar and M. Suchetha, "An approach for analysis and prediction of CKD using deep learning architecture," in *Proc. Int. Conf. Commun. Electron. Syst. (ICCES)*, Jul. 2019, pp. 1660–1664.
- [35] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using Twitter, apache spark, and machine learning," *Appl. Sci.*, vol. 10, no. 4, p. 1398, Feb. 2020.
- [36] T. Muhammed, R. Mehmood, A. Albeshri, and I. Katib, "UbeHealth: A personalized ubiquitous cloud and edge-enabled networked healthcare system for smart cities," *IEEE Access*, vol. 6, pp. 32258–32285, 2018.
- [37] P. M. Kumar and U. Devi Gandhi, "A novel three-tier Internet of things architecture with machine learning algorithm for early detection of heart diseases," *Comput. Electr. Eng.*, vol. 65, pp. 222–235, Jan. 2018.
- [38] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *Proc. 4th Int. Conf. Cyber IT Service Manage.*, Apr. 2016, pp. 1–6.
- [39] S. Pai, S. Hui, R. Isserlin, M. A. Shah, H. Kaka, and G. D. Bader, "NetDx: Interpretable patient classification using integrated patient similarity networks," *Mol. Syst. Biol.*, vol. 15, no. 3, p. e8497, Mar. 2019.
- [40] J. Chen, L. Sun, C. Guo, W. Wei, and Y. Xie, "A data-driven framework of typical treatment process extraction and evaluation," *J. Biomed. Informat.*, vol. 83, pp. 178–195, Jul. 2018.
- [41] X. Zeng, Z. Jia, Z. He, W. Chen, X. Lu, H. Duan, and H. Li, "Measure clinical drug–drug similarity using electronic medical records," *Int. J. Med. Inform.*, vol. 124, pp. 97–103, Apr. 2019.
- [42] M. Norouzi, D. J. Fleet, and R. Salakhutdinov, "Hamming distance metric learning," in *Proc. NIPS*, 2012, pp. 1070–1078.
- [43] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang, "Measuring patient similarities via a deep architecture with medical concept embedding," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 749–758.
- [44] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE Trans. Nanobiosci.*, vol. 17, no. 3, pp. 219–227, Jul. 2018.
- [45] F. Wang and J. Sun, "PSF: A unified patient similarity evaluation framework through metric learning with weak supervision," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 3, pp. 1053–1060, May 2015.
- [46] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang, "Metric learning on healthcare data with incomplete modalities," *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3534–3540.
- [47] N. Dohare and S. Sachdeva, "Evaluation of nephrology dataset through deep learning technique," in *Communications in Computer and Information Science*, vol. 1229. Singapore: Springer, 2020, pp. 131–139.
- [48] Government of India. (2019). *Placing the Report on National Digital Health Blueprint (NDHB) in Public Domain for Comments/Views Regarding*. [Online]. Available: https://mohfw.gov.in/sites/default/files/National_Digital_Health_Blueprint_Report_comments_invited.pdf
- [49] D. Teodoro, E. Sundvall, M. João Junior, P. Ruch, and S. Miranda Freire, "ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers," *PLoS ONE*, vol. 13, no. 1, Jan. 2018, Art. no. e0190028.
- [50] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "–Omic and electronic health record big data analytics for precision medicine," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 263–273, Feb. 2017.
- [51] J. L. Schafer, "Multiple imputation: A primer," *Stat. Methods Med. Res.*, vol. 8, no. 1, pp. 3–15, Feb. 1999.
- [52] K. A. Hallgren and K. Witkiewitz, "Missing data in alcohol clinical trials: A comparison of methods," *Alcoholism: Clin. Exp. Res.*, vol. 37, no. 12, pp. 2152–2160, Dec. 2013.
- [53] J. Tian, B. Yu, D. Yu, and S. Ma, "Missing data analyses: A hybrid multiple imputation algorithm using gray system theory and entropy based on clustering," *Int. J. Speech Technol.*, vol. 40, no. 2, pp. 376–388, Mar. 2014.
- [54] N. Stadler, "Pattern alternating maximization algorithm for missing data in high-dimensional problems," *J. Mach. Learn. Res.*, vol. 15, pp. 1903–1928, Jan. 2014.
- [55] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.
- [56] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton, "On rectified linear units for speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3517–3521.
- [57] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2014, pp. 165–180.
- [58] J. Bromley, I. Guyon, Y. LeCun, E. Sckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6, 1994, pp. 737–744.
- [59] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," Tech. Rep., 2016. [Online]. Available: <https://arxiv.org/abs/1603.04467>
- [60] F. Chollet. (2015). *Keras*. [Online]. Available: <https://github.com/fchollet/keras>
- [61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.
- [62] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.
- [63] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. JMLR Workshops Conf.*, vol. 27, 2012, pp. 17–36.

- [64] V. Gupta, S. Sachdeva, and N. Dohare, "Deep similarity learning for disease prediction," in *Book Trends in Deep Learning Methodologies—Algorithms, Applications, and Systems* (Series Hybrid Computational Intelligence for Pattern Analysis and Understanding). New York, NY, USA: Academic, Dec. 2020.
- [65] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011.



SHELLY SACHDEVA received the B.E. and M.Tech. degrees (Hons.) in computer science, in India, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from The University of Aizu, Japan, in 2012.

She is currently working as an Associate Professor and the Head of the Department of Computer Science, National Institute of Technology (NIT), Delhi, India. She is having more than 15 years of experience in the area of teaching and research. She has published more than 60 papers in reputed/ refereed journals and conferences. Her main research interests include database systems, the high-level query interfaces, security in databases, computational intelligence, interoperable electronic health records, and data quality for health informatics.

Dr. Sachdeva received the Japanese MEXT Scholarship from April 2009 to March 2012. She was awarded the Gold Medal for Masters of Technology in computer science in 2004. She achieved the Indian Government Scholarship for Master’s Program in 2002. She qualified the Graduate Aptitude Test in Engineering (GATE) in 2002. She attended many conferences, workshops, and presented posters/papers. She organized workshops, faculty development programs, and was a Coordinator of master program at IIIT. She is a Reviewer and a member of several transactions and international conferences, including IEEE.



VAGISHA GUPTA received the B.Tech. degree in computer science and engineering from Shri Mata Vaishno Devi University, Jammu and Kashmir, India, in 2017. She is currently pursuing the M.Tech. degree in computer science and engineering with specialization of data analytics with the National Institute of Technology Delhi, Delhi.

She is also involved in contributing to open source projects and worked with Open Information Security Foundation as an Outreachy Intern.

Her research interests include databases, deep learning, data science and analytics, data driven programming, and information security.

Ms. Gupta achieved the Indian Government Scholarship for Master’s Program for the session 2018–2020. She qualified the Graduate Aptitude Test in Engineering (GATE) in 2018. She organized and participated in various workshops and events during her bachelor’s program at SMVDU.



SUBHASH BHALLA (Member, IEEE) received the B.Tech. degree, the master’s degree in computer science, and the Ph.D. degree from IIT Delhi.

He joined the School of Computer and Systems Sciences (SCSSS), Jawaharlal Nehru University (JNU), New Delhi, as a Faculty Member, in 1986. He was a Visiting Scientist at the Sloan School of Management, Massachusetts Institute of Technology (MIT), Cambridge, MA, from 1987 to 1988. He is currently with the Department of Computer Software, The University of Aizu. He is also exploring database designs to support models for information interchange through the World Wide Web. He is a member of a Study Team on recovery and transaction management system for mobile computing. He is studying transaction management and algorithmic designs for distributed real-time systems. His research interests include performance evaluation and modeling of distributed algorithms, managing components and application services, and integration of technologies. He is a member of the IEEE Computer Society and SIGMOD of ACM.

...