A Comprehensive Review of Methods Used for Health Prediction and Monitoring Utilizing an Electronic Medical Records (EMR) System

SP Jayasekera[#], LP Kalansooriya
Department of Computer Science, Faculty of Computing Sir John Kotelawala Defence University, Sri Lanka
#37-cs-0009@kdu.ac.lk

ABSTRACT: In the rapidly evolving field of healthcare, Artificial Intelligence (AI) and pattern recognition play a key role in enhancing disease diagnosis and prediction. As the patient population increases, the digitalization of medical records has become essential, therefore electronic medical records were developed. This stored Electronic Medical Records (EMR) data can be used to predict possible diseases based on the symptoms stored in the system. This study delves into the integration of AI methodologies within EMR systems, providing a comprehensive review of current techniques that have been used in health prediction and monitoring using EMR data. In this paper, different AI-driven approaches were examined and compared, including Deep Learning (DL), Machine Learning (ML), and Rule-Based Methods. This paper reveals the potential of these techniques in accurately diagnosing diseases, additionally, it discusses challenges and future directions, emphasizing the need for innovative solutions to optimize EMR systems in the context of AI and pattern recognition. Several instances where AI models, such as the application of Support Vector Machine (SVM) models, achieved predictive accuracies of 86.2% and 97.33% in different cancer types, and ML models diagnosing Diabetic Retinopathy with a 92% accuracy rate were observed. Variations in the effectiveness of these technologies across different diseases were also observed, such that a technique that has high accuracy in one disease may have lower accuracy in a different disease. This paper aims to contribute to the growing body of knowledge in AI applications in healthcare, offering insights into the development of more efficient, accurate, and predictive healthcare models.

INDEX TERMS: Healthcare, Deep learning, Electronic Medical Records, Rule-based method, Disease diagnosis, Machine learning.

I. INTRODUCTION

One of the most critical responsibilities of medical institutions is managing patient data, and a patient file is an essential source of data since it enables the development of comprehensive healthcare strategies. It had been common practice for a long time to keep records on paper where medical offices, hospitals, and clinics frequently gathered files and kept patient history using a paper record system. However, paper medical records have a lot of drawbacks such as insufficient storage space, insufficient backups, inconsistency in the layout, and unclear audit trails. Due to technological advancements, electronic medical records were introduced to store patient data on computers or smart devices and overcome paper records' drawbacks.

Electronic Medical Records (EMR) are digitalized versions of paper charts in clinics and hospitals. Clinicians and doctors primarily use these EMRs to diagnose and treat patients and record information by and for the physicians in the hospital. The use of EMRs has become increasingly prevalent in healthcare, with potential benefits such as improved patient care and reduced medical errors [5]. It contains a patient's medical history, diagnoses, prescriptions, treatment schedules, vaccination dates, and lab and test results. These are stored in databases that enable doctors or clinicians to access patient information quickly, track vaccinations, follow patient health performance, and make informed judgments with proper understanding and confidence for the most complex multi-axial diseases, heart diseases, and cancers [4].

By computerizing patient information, there is also a significant change in how patient data are arranged and made available for applications that weren't previously possible with paper records. Thus, it shows that the main objective of an EMR is keeping an eye on the patient while improving healthcare quality. It is important to understand the patient's unique perspective and experiences in the diagnosis and treatment of disease, using EMR has been shown to improve patient outcomes and satisfaction, as well as enhance the physician-patient relationship [2]. Even though EMR provides users and physicians with several advantages, several difficulties are connected to their implementation, such as computer downtime, computer professionals' limitations, a lack of user communication, security risks of confidentiality-leakage, etc which should be considered [6]. An accurate and timely diagnosis is the foundation of any successful treatment. Access to longitudinal data from a patient's EMR might be a valuable clinical resource that could be utilized to forecast

future events or diagnoses [1]. A patient's status is thoroughly described in an EMR, and applying data-driven technologies to an EMR enables us to accurately predict and diagnose diseases. This can be made possible by making the raw EMR data into a machine learning representation or turning the data into relevant data that can be processed algorithmically. The integration of AI technologies with EMR systems represents a groundbreaking development in this context. AI's ability to process large datasets and uncover patterns offers unparalleled opportunities for improving disease diagnosis, treatment planning, and patient monitoring.

There are different types of data-driven techniques used to accomplish prediction and diagnosis systems that medical professionals can employ to effectively forecast illnesses and enhance the health of their patients. This review aims to find the most accurate methods for diagnosing and predicting diseases by describing and comparing various methods and techniques used for health prediction and monitoring using EMR.

This study discusses numerous disease diagnosis and prediction methods using electronic records, highlighting their benefits and drawbacks. It also discusses current trends and potential future developments and makes a comparative comparison of the various methods.

The literature review of this paper explores the significance of EMR data in monitoring patient health and advancing data-driven decision-making. It delves into the growing interest in employing computer-assisted methods for disease diagnostics based on Electronic Health Record (HER) data, categorizing these methods into distinct approaches. Machine Learning (ML) methods, encompassing Bayesian, Support Vector Machine (SVM), and decision tree techniques, are discussed, along with the challenges of integrating raw EHR data into ML models due to complexity and limited healthcare data. Bayesian Networks are highlighted for their use in probabilistic medical ontology reasoning, aiding in disease diagnosis and prediction. Decision Trees are emphasized for their effectiveness in the early identification of diseases like Diabetic Retinopathy and asthma. Additionally, rule-based heuristic techniques are explored for diagnosing colorectal cancer and lupus. Finally, Deep Learning methods, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Deep Belief Networks (DBN), and Autoencoders (AE). Using these findings, it is aimed to present a comprehensive overview of the existing predicting systems implemented using the above-mentioned techniques and EMR data.

The paper is structured into five sections. Section 2 discusses the current research on methods for disease diagnosis. Section 3 is the Methodology. Section 4 contains the discussion. Finally, Section 5 presents the conclusion of the review.

II. DISEASE DIAGNOSIS USING DATA-DRIVEN MODELS

EMR data is a critical resource in modern healthcare, providing a dynamic method to monitor patient health and improve decision-making using data-driven solutions. Unlike traditional clinical tests and biological investigations, the fundamental goal of EMR data is to track a patient's health over time in a methodical manner. This large set of patient data has paved the way for the creation of prediction models by implementing AI models such as Machine Learning (ML), Deep Learning (DL), and Rule-Based Methods, which have revolutionized disease prediction and diagnosis processes.

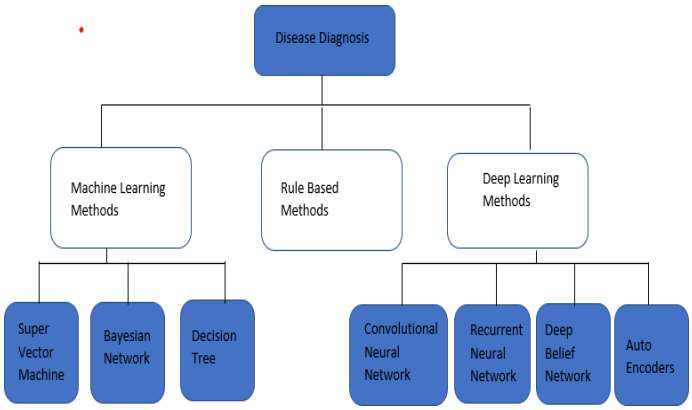

Figure 01. Breakdown of the techniques used in this review to diagnose diseases.

This paper discusses various electronic medical record-based methods for diagnosing diseases automatically. Depending on their technique, models have been grouped into different approaches to diagnosing diseases using EMR data.

A. Machine Learning (ML) Methods

Health database systems based on electronic medical records (EMR) are most often created using machine learning methods for individuals who have had health examinations [7]. Machine learning methods can be categorized into different approaches, including Bayesian, SVM, and decision tree methods. Each of these approaches represents a distinct category within the field of machine learning [34].

Many research studies have used EHR data for a predictive model, which involves constructing a statistical model to predict a clinical outcome using machine learning. However, it is difficult to directly integrate raw EHR data into ML models for predictive models due to the complexity of EHR data [8]. Because a lack of data prevents machine learning from solving many healthcare issues.

1) Support Vector Machine (SVM)

For supervised classification, experts often use Support Vector Machines (SVM). SVM is based on labeled data, and Vapnik invented SVM [45]. A training dataset is used to find data from

the input that has a structure like the output data when both the input and the output have already been supplied.

Getting a cancer diagnosis is crucial for prospective patients since early tumour identification and therapy can improve survival. In [9], a cancer diagnosis was performed using the SVM model using medical information retrieved straight from the EHR. As part of the proposed approach, SVM models for cancer classification were trained using medical records extracted from Electronic Health Records (EHRs). These SVM models Based on the medical data that was analysed, played a crucial part in the cancer categorization procedure. After being trained on 400 pieces of data for each cancer and employing 100 pieces of health information for each cancer, the algorithm has shown a predictive accuracy of 86.2% for ten different forms of cancer and 97.33% for three different types of cancer.

An SVM-based technique was used [10] for significant cohort research to diagnose contralateral breast cancer. Characteristics based on pathology reports for every area of breast cancer and narrative text in progress notes were used to derive features, Zeng et al. [10] designed and put into practice a novel methodology. The suggested strategy for identifying contralateral occurrences in the notes uses medical ideas and how they are combined. SVM and derived characteristics are used to detect contralateral cancers. During the validation process, the area under the curve (AUC) for the model was determined to be 0.93, indicating its high accuracy in predicting outcomes. In the test set, the AUC was slightly lower at 0.89, indicating a slightly reduced but still reliable performance. This strategy of feature development is advantageous due to its simplicity and can be applied to different occurrences of breast cancer as well as to identify various other diseases.

To identify Rheumatoid Arthritis (RA) patients, the Support Vector Machine (SVM) technique can be employed. This technique utilizes a set of naïve and expert-defined Electronic Health Record (EHR) characteristics for the identification process [12]. This method uses Natural language processing (NLP) concepts, pharmaceutical exposures, and billing codes. The SVM methodology was trained using both expert-defined and naive data. The accuracy and recall scores were 0.94 and 0.87, respectively, as opposed to 0.75 and 0.51 for deterministic approaches. In this study, a dataset of 10,000 patients was employed. The test findings divided the patients into three groups: potential RA, definite RA, and not RA.

2) Bayesian Network (BN)

A probabilistic graphical framework called a Bayesian network is utilized to represent a group of variables and their conditional interactions. This graphical model employs a directed acyclic graph (DAG) to illustrate the relationships among the variables and their dependencies. Naive Bayes (NB) and Bayesian Networks (BN) are both probabilistic algorithms that perform effectively with various characteristics [14].

Building Clinical Bayesian Networks (CBN) for probabilistic medical ontologies reasoning is described in [13] to directly learn the entire ontology and high-quality Bayesian topology from EMRs. More than 10,000 patient records analysed for medical entity connections have used the K2 greedy method and Odds Ratio (OR value) computation to create a Bayesian topology automatically. The study demonstrates that medical information can generate high-quality health topology and ontology directly and automatically. A clinical Bayesian network has been developed using the study's probability distribution between illness and other parameters. With 1712 test samples, an accuracy of 64.83% was produced by the Naïve Bayesian network, while the Basic Bayesian network produced 68.45%.

In a study by Sakai et al. [15], they evaluated the diagnostic performance of a Bayesian network in comparison to the NB model, an artificial neural network (ANN), and a logistic regression model to identify instances of appendicitis. 169 people who were thought to have acute appendicitis were included in the dataset for the study. The performance of the proposed model was assessed using logistic regression and neural network metrics. Compared to other diagnostic models examined in this research, this model had the lowest error rate and produced the most trustworthy findings, detecting that 50.9% of patients (86 out of 169) had appendicitis.

The Naïve Bayes method was employed in Al-Aidaroos et al. [16] review of medical data mining to classify medical data and diagnoses such as primary tumours, hepatic issues, and breast or lung cancer. Using 15 datasets, the proposed NB strategy was empirically compared with five other approaches to show its superiority. The findings indicated that NB performed better than others regarding medical categorization. Deep learning ideas can produce superior segmentation results with the proposed approach. The report states that future research will combine NB and different methodologies.

Kazmierska and Malicki researched the Bayesian classifier, which is used to assess whether cancer is progressing or relapsing [17]. This study analysed data from 142 individuals who had radiation therapy for brain tumours between 2000 and 2005. For training, 96 binary attributes were selected. As a result of the proposed model, the likelihood of having a cancer relapse has been determined as well as the likelihood of not having one. The proposed method received scores of 0.84, 0.87, and 0.80 for accuracy, specificity, and sensitivity, respectively.

3) Decision Tree

EHR data can accelerate and simplify the early identification of Diabetic Retinopathy (DR). Five machine-learning techniques are used in [18] to identify diabetic retinopathy using electronic health record data. Records from 301 Chinese hospitals were compiled into a sizable retinal dataset. To increase the accuracy of DR illness diagnosis, preprocessing techniques such as label binarization, value normalization, and standard acceleration are carried out. According to the experimental findings, the machine learning model's Random Forest (RF) can achieve an accuracy level of 92% while performing well. Due to its low cost, low threshold, and

excellent diagnosis accuracy, the suggested approach has an advantage over current DR diagnostic methods.

The primary objective of the study conducted by Lungu et al. [19] was to investigate whether machine learning techniques could enhance the diagnostic precision of Magnetic Resonance Imaging (MRI) in detecting pulmonary hypertension (PH). This was accomplished by employing computational modelling approaches and image-based metrics. MRI as well as the Right Heart Catheterization (RHC) were used to identify PH using a decision tree method [19]. Seventy-two individuals with potential PH underwent MRI and RHC, and 57 of these patients were found to have the condition, while 15 samples were determined to be PH-free. As a result of the proposed algorithm, 92% of the PH cases were correctly identified, while 4% were misclassified. If the findings of this study are as anticipated, RHC may not be required when PH is suspected.

In [20], the decision tree is used in the first phase to diagnose asthma, and the fuzzy system is utilized in the second phase to assess the level of asthma management. Dry cough, sore throat, sneezing, and other symptoms have been used to diagnose asthma, whereas breathlessness and other daytime symptoms have been used to measure the control level. In this study, the information was gathered through the patients' responses to questionnaires. Diagnoses of asthmatic patients were made using a decision tree classifier, which had accuracy and kappa coefficients of 0.90 and 0.783, respectively.

B. Rule-Based Method

In [22], the diagnosis of colorectal cancer was made using a rule-based heuristic technique. Machine learning and rule-based methods' effectiveness was evaluated for each phase. The algorithm identified concepts at the document level with an F-measure of 0.996 as well as detected cases at the patient level with 0.93 for the F-measure using the manually examined data set of 300 potential Colorectal cancer patients. In the work by Breischneider et al. [23], in this study, rule-based grammar was used to obtain textual information from records of patients with mamma carcinoma. Based on recovered textual fragments, seven essential criteria were listed to construct the therapeutic suggestion. The mammography use case was used to assess the proposed system. With an accuracy of 0.69, a textual feature extraction approach based on rule-based decision support, information extraction, and semantic modelling was employed to determine the lymph node status.

In an EHR dataset with 400 records, Jorge et al. [24] used rule-based approaches to identify lupus patients. Natural language processing was used to extract the narrative and codified data from the training set of data (NLP). Based on penalized logistic regression, the author classified systemic lupus erythematosus (SLE) as either definite or probable. The machine learning code utilized in this work for definite SLE showed a 90% positive predictive value, with a specificity of 97%. According to the best rule-based method (ICD-9 code), the specificity and sensitivity were respectively 86% and 84 % and 60 % and 69 % for definite and definite/probable SLE.

C. Deep Learning Methods

Deep neural networks, including autoencoders (AE), Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), Recurrent Neural Networks (RNN), and other similar architectures, are considered the most effective machine learning techniques in the biomedical sector [25]. These networks form the foundation of deep learning and have shown remarkable effectiveness in various biomedical applications. Various deep learning methods used on electronic medical records are examined in this review to apply them to clinical tasks. Their benefits are discussed in practice and potential future applications.

1) Convolutional Neural Network (CNN)

A method for unsupervised deep feature learning that Miotto et al. introduced in [21]. Using clinical notes as the input, they drove patient representation in their predictive modelling technique. By identifying hierarchical regularities and relationships in clinical notes, 700,000 individuals from the Mount Sinai dataset were used. The study encompassed a broad range of clinical areas and chronological periods, involving a total of 76,214 test individuals, representing 78 distinct diseases. The study's findings surpassed approaches that relied on a representation derived from basic medical information, where accurate and F-score forecasts improved by 92.9% and 18.1%, respectively. When produced patient representations are included in DL approaches, clinical prediction can be improved. This study can use the laboratory findings to improve the quality of its model.

Multiple illnesses have been evaluated using the disease prediction model built on EMRs [26]. The Convolutional Neural Network (CNN) has been used to characterize the suggested strategy for multiple illness prediction. This approach was tested on 4298 patients with a brain infection, coronary heart disease, and pulmonary infection. In a dataset for cerebral infections, the CNN algorithm, the accuracy was 96.5% and the F1-measure score was 96.6%.

2) Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNN) were specifically designed to process sequential inputs, such as language data. The present state of an RNN implicitly incorporates knowledge about the whole history of the series since RNNs process, a series of inputs that transmits the concealed value of every input unit to the next input unit, one item at a time. Doctor AI [27], which was over eight years, performed over 260K time-stamped analyses on individuals' electronic health records longitudinally, which is one RNN-inspiring technique. Doctor AI surpassed multiple baselines and scored 79.58% on a sizable real-world EHR dataset.

Wu et al. [28] presented a novel approach for categorizing paediatric asthma by utilizing event sequences and their corresponding characteristics. The findings of this study show that including a timestamp in an RNN model enhances the

categorization of individuals without asthma rather than those who have it.

3) Deep Belief Network (DBN)

A Deep Belief Network (DBN) has been used to diagnose Parkinson's disease (PD) using speech sounds collected from the UCI repository [30]. A range of healthy and sick voices was used to train the suggested approach, using DBN as a data source, and the features were extracted. According to the proposed method, the PD consists of one output layer and two stacked limited Boltzmann machines. Parkinson's disease was diagnosed with 94% accuracy using the recommended approach.

DBN has been used [31] to diagnose attention deficit hyperactivity disorder (ADHD), which is one of the most common diseases. The network was built and trained using a greedy methodology according to the recommended strategy. The Global Competitions ADHD-200 has provided the two training and testing datasets. This study has used samples from the Neuroimaging (NI) and New York University (NYU) databases for training and testing, respectively. These findings show that they attain cutting-edge accuracy of 0.6368 on the NYU dataset and 0.6983 on the NI dataset.

4) Auto Encoders (AE)

In a study, researchers employed auto-encoders to forecast a particular group of diagnoses [29]. For the detection and classification of softmax, stacked autoencoder, and cervical cancer classification algorithms have been utilized [32]. To train and test the approach, the UCI dataset with 30 characteristics, four targets, and 668 samples was used. A training set made up 70% of the dataset, while a test set made up 30%. Four target variables were applied to the suggested model, and the efficacy of its categorization was evaluated. This comparison produced a 0.978 accurate classification rate. Due to the dimensionality reduction of the samples, this model's training takes far too much time. In the future, advanced methods could be used to reduce the training time of the model. Hwang et al. [33] examined the efficacy of missing value prediction, conventional networks, and generative adversarial networks (GANs) methods combined for illness prediction [33]. With a specificity of 0.99, along with a sensitivity of 0.95, also with an accuracy of 0.98, the stacked autoencoder (missing value forecasting technique) and auxiliary classifier GANs (AC-GANs: illness prediction) have shown excellent results. In this work, AE fills in the gaps left by the GAN generic model. The use of GAN to fill in the missing data is one of this work's future directions.

III. METHODOLOGY

An efficient and effective way to obtain requirements is through document analysis, which involves reviewing current system documentation and acquiring data. In the study, a systematic analysis of 40 papers was conducted, and 31 of them were selected based on stringent criteria to review in this paper. Research articles were used from Google Scholar and other research archives articles on EMR-based disease diagnosis based on AI methods such as ML, Rule-based, and DL methods. The selection process involved multiple stages, including title and abstract screening, full-text reviews, and a backward and forward search to capture additional relevant works. Selected literature used specific keywords and phrases, and combinations of these terms. Keywords related to the topic were searched to find existing research articles.

Several research articles on EMR systems and methods used for predicting systems were reviewed and analyzed. The research article categorizes the methods used to track and predict health into three categories: Different approaches were employed in the study, including the utilization of Rule-Based Methods, Machine Learning (ML) Methods, and Deep Learning (DL) Methods. These categories were chosen based on the predominant analytical techniques used in the articles and provided a structured way to compare the effectiveness and accuracy of different methods.

To simplify data analysis, the literature review was summarized into tables. Tables provide an overview of methods used, the disease addressed in the paper, performance measures such as accuracy and F-measures (a measure of test accuracy), and the paper's objectives. Comparing the effectiveness and accuracy of different methods can be done using the tables.

Support Vector Machines (SVMs) are strong technology that includes both nonlinear and linear regression approaches, making them essential to data mining processes. SVMs can conduct multiclass and binary classification, making them useful for data prediction and classification, including in the field of health research. SVMs are frequently used by researchers for supervised classification, particularly in disease detection and prediction.

For instance, SVM methods have been used to identify different types of diseases, such as breast cancer and rheumatoid arthritis, with high accuracy. Zhang et al. [9] achieved 97.33% accuracy in cancer classification from Electronic Health Records (EHRs) using SVM, while Zeng et al. [10] conducted a validation study on detecting contralateral breast cancer, achieving a high area under the ROC curve (AUC) of 93% when utilizing extracted features in combination with pathology reports. Additionally, SVMs have been employed in the identification of rheumatoid arthritis (RA) phenotypes, achieving an F-Measure score of 88.6% in a Naïve EHR with a sample size of 376 patients [12].

Table 6. The Summary of the SVM Methods

Methods	Focused	Performance	Dataset	Objective

	disease	Measures		
SVM- RBF [9]	Cancer	Accuracy- 97.33%	Employed 100 pieces of health information for each cancer and trained on 400 pieces of data for each cancer.	Using SVMs to classify cancer from EHRs.
SVM [10]	Breast Cancer	Testing- 89% AUC Validation- 93%	A total of 1063 women with breast cancer.	Analyzing pathology reports and extracted features to identify contralateral 1 breast cancer.
SVM [12]	Rheumatoid Arthritis	Precision- 96.8% F-Measure- 88.6% Recall- 87% AUC-96.6%	In total, 376 patients (185 with RA and 191, not RA).	SVM-based phenotyping of RA in the Naïve EHR.

The below graph shows the average performance based on the performance measure obtained for the different diseases based on the SVM methods used.

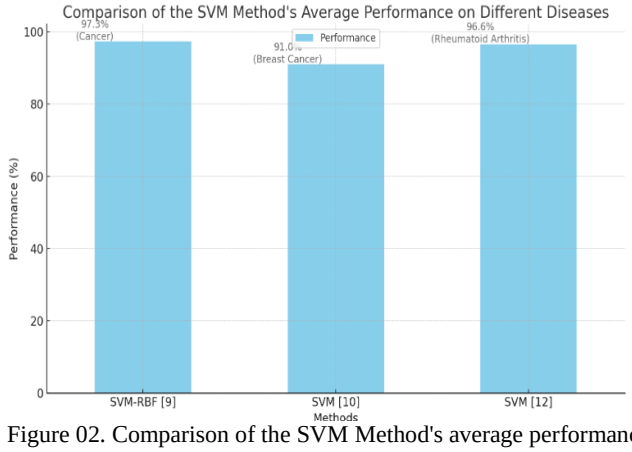

Figure 02. Comparison of the SVM Method's average performance on different diseases

Bayesian approaches, such as Naive Bayes (NB) and Bayesian Networks (BN), are probabilistic algorithms that use multiple features elegantly. By utilizing proven biomarker operating characteristics, Bayesian clustering can accommodate patients with varied data availability. The advantage of Bayesian joint modeling is that it incorporates phenotypic uncertainty into future association analyses, producing correct uncertainty estimates. When compared to Bayesian networks, NB classifiers do not require dependency networks and are better at handling high-dimensional features. This research looks at four articles that utilize Bayesian approaches to predict diseases like cancer, appendicitis, hepatitis, and brain tumors.

Shen et al. [13] tested the accuracy of Naive Bayes and Bayesian Networks in predicting cancer and achieved 64.83% and 64.83% accuracy, respectively. Sakai et al. [15] used the Bayesian network to predict the diagnosis of acute appendicitis. Aidaroos et al. [16] classified cancer, hepatitis, and

liver disorders using NB with an accuracy of 97.43%. Bayesian networks were also used to optimize treatment decisions for a brain tumor with 84% accuracy. The table below lists a few Bayesian method-based systems, and research articles are used to note how accurate the results were when used to predict diseases.

Bayesian statistical models have been utilized when there are gaps in the information provided by local data but there are additional sources of information that can help close the gaps. There are further advantages to Bayesian modeling since it gives a reasonable framework for incorporating new data as it becomes available and helps practitioners to rapidly estimate future illness scenarios. The table below lists a few Bayesian method-based systems, and research articles are used to note how accurate the results were when using Bayesian method-based to predict diseases.

Table 7. The Summary of the Bayesian Methods

Methods	Focused Disease(s)	Performance Measures	Dataset	Objective
NB, BN [13]	cancer	NB Accuracy- 64.83% BN Accuracy- 68.45%	Records of 10,000 identified patients.	An Automatic Bayesian topology generation using the K2 greedy method and odds ratios (OR values).
Bayesian Network [15]	Appendicitis	-	A database contains 169 people who may have acute appendicitis.	An algorithm for predicting acute appendicitis using Bayesian networks.
NB, LR, DT, and NN [16]	Multiple diseases, including cancer, hepatitis, and liver disorders	Accuracy- 97.43% AUC- 99%	Various illnesses are illustrated in 15 datasets from the UCI library.	LR, NB, NN, and DT classification of medical data.
Bayesian Network [17]	Brain Tumor	The accuracy rate is 84% The sensitivity is 80% The specificity is 87%	142 patients with brain tumors.	Optimization of treatment decisions using the Naïve Bayesian Classifier.

The graph illustrates the average performance of different diseases based on the use of Bayesian methods for diagnosis.

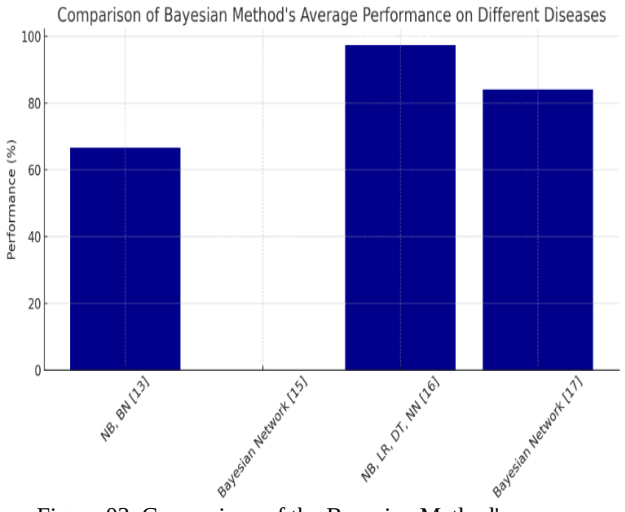

Figure 03. Comparison of the Bayesian Method's average performance on different diseases

Using decision trees was another method used for predicting diseases in research articles. A decision tree aids in creating a fair picture of the rewards and hazards related to each potential result. When contemplating EHRs, where uncertainty is prevalent, decision trees are highly helpful because they are especially beneficial when the results are unknown. A decision tree is an effective tool for decision-making. It offers a useful framework within which to consider options and investigate what might result from each.

Decision trees are used to categorize records, which are useful for challenges involving association and regression. By using a decision tree, advantages and disadvantages can be quickly visualized and identified. The diagnosis system for diabetic retinopathy developed by Sun and Zhang [18] achieved 86.82% accuracy. Based on MRI images, Lung et al. [19] were able to diagnose pulmonary hypertension with 92% accuracy using a decision tree. Using a decision tree and fuzzy system, asthma diagnosis and control levels were
determined [20].

The reliability and effectiveness of decision trees in medical decision-making are supported by reputable sources, including research articles and academic publications. Decision trees provide high classification accuracy and are a dependable and effective means of making judgments due to their plain representation of the information gathered. They have been widely used in a variety of medical decision-making scenarios, including classification and diagnosis. The fundamental properties of decision trees and their effective applications in medicine have been emphasized in the literature, highlighting their potential for future use in medical research and practice.

Table 8. The Summary of the Decision Tree Methods

Methods	**Focused Disease(s)**	**Performance Measures**	**Dataset**	**Objective**
DecisionTree [18]	Diabetic Retinopathy	Accuracy- 86.6%	301 Chinese hospitals provided 5057 records.	Five machine learning techniques are used with the EHR to diagnose DR.
DecisionTree [19]	Pulmonary hypertension	Sensitivity is 97% Accuracy of 92% Specificity- 73%	Pulmonary hypertension is suspected in 72 patients.	Analyzing MRI images to diagnose pulmonary hypertension.
Decision Tree and Fuzzy system [20]	Asthma	Kappa- 78.32% Accuracy- 90%	30 of patients with asthma.	Using fuzzy logic and decision trees to diagnose and control asthma.

The graph below presents the average performance of various diseases using Decision tree methods for diagnosis.

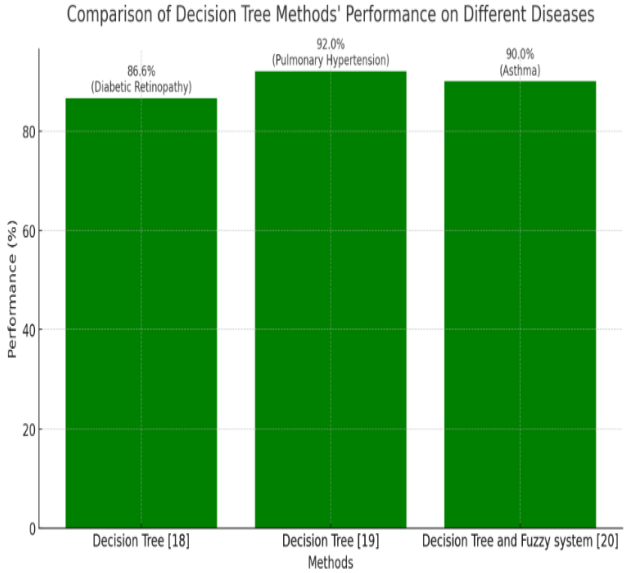

Figure 04.Comparison of the Decision Tree Method's average performance on different diseases

By using rule-based systems, we can retrieve features from electronic medical records quickly. For the extraction of data, rule-based systems are used since the most common kind of knowledge representation is if-then logic.

Using rule-based systems, domain experts can express and rate their expertise. The decision-making process can then use that data. To determine the outcomes of rule-based or identically based systems, users must input specific attributes or facts, such as patient symptoms. It is difficult for someone without medical training to do this. A drawback of this method is the

requirement for precise definitions of data properties. Using the rule-based method, computer scientists identify rules and identify patterns associated with them. Xu et al. [22] used this method to identify colorectal cancer. Breischneider et al. [23] used automated breast cancer detection using rule-based grammar and achieved 90% accuracy. Using a rule-based algorithm and machine learning codified algorithm, Jorge et al. [24] identified Lupus patients from EMR.

Table 9. The Summary of the Rule-based Methods

Methods	Focused Disease(s)	Performance Measures	Dataset	Objective
Rule-based ML-based [22]	Colorectal cancer	Accuracy- 99.6% Precision- 99.6% Specificity- 96.9% F-measure- 99.6%	1,262,671 patients from a synthetic derivative database.	Data extraction and integration from EHRs for Colorectal cancer detection
Rule-based grammar approach [23]	Breast cancer	Accuracy of 90% The Specificity of 59% Sensitivity of 98%	The university hospital in Erlangen collected the clinical reports of 2096 patients totaling 8766.	Clinical report information extraction for breast cancer.
The rule-based algorithm is, Machine learning codified algorithm. [24]	Definite SLE Definite probable SLE	Sensitivity- 86% Specificity- 60% PPV- 46% Sensitivity- 84% Specificity- 69% PPV 65%	400 records in an EHR dataset.	From the EHR, recognize patients with Lupus.

The graph illustrates the performance of rule-based methods applied to the diagnosis of various diseases.

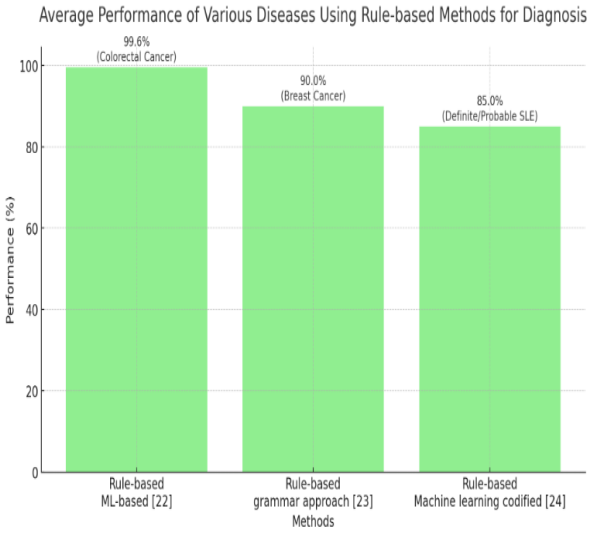

Figure 05. Comparison of the Decision Tree Method's average performance on different diseases

Deep Learning (DL), often referred to as hierarchical learning, is a sophisticated modeling approach characterized by its use of multiple processing layers to analyze complex data sets. This method is increasingly employed in the analysis of the ever-expanding volumes of EHRs. The application of deep learning in the realm of EHRs is particularly notable in research endeavors focused on forecasting individual health outcomes and assessing potential risks. At the heart of deep learning technology are various types of neural networks, each with unique capabilities and applications. These include convolutional neural networks (CNNs), known for their prowess in processing visual imagery; recurrent neural networks (RNNs), which excel in handling sequential data; deep belief networks (DBNs), which are effective in probability-based learning; and autoencoders, specialized in data encoding and reconstruction tasks. These diverse neural network architectures enable deep learning to effectively interpret and utilize the vast and complex data present in EHRs for advanced medical research and analysis.

Table 5. The Summary of the Deep Learning Me

Methods	**Focused Disease(s)**	**Performance Measures**	**Dataset**	**Objectives**
Unsupervised deep feature learning [21]	78 diseases	Accuracy- 92.9% F-score- 18.1%	The Data warehouse from Mount Sinai contains 700,00 patients.	Predictive models can be developed using patient representations from EHRs.
CNN and Framingham risk score [26]	Cerebral infraction (CI), Pulmonary Infarction (PI), And Coronary Heart (CH)	Accuracy CI-96.5% PI- 95.6% CH- 93.6%	From a Chinese hospital with a grade-A rating, 4298 individuals were evaluated.	Clinical notes based on a uniform model for assessing multiple diseases.
RNN [27]	Numerous diseases	Recall- 79.58%	260K patients.	Applied to longitudinally timestamped EHRs.
RNN [28]	Pediatric Asthma	Precision- 84.54% F-measure- 85.08% Recall- 85.65%	4000 patients from Physionet and 4013 patients from Olmsted Country Birth Cohort.	RNN-based asthma classification in pediatrics.
DBN [30]	Parkinson's Disease	Accuracy- 94%	Data set on 31 Parkinson's patients.	DBN-based Parkinson's disease diagnosis system.
DBN with greedy Approach [31]	ADHD	NI- 69.83% Accuracy- NYU- 63.68%	Neuroimaging-samples of 73 New York University-samples of 263.	A greedy approach to the diagnosis of ADHD using DBN.
Stacked AE and Softmax classification. [32]	Cervical cancer	Accuracy- 97.25%	668 samples from the UCI dataset.	Stack autoencoder and softmax classification for cervical cancer classification and diagnosis.
Stacked AE and GAN [33]	Breast cancer	Sensitivity- 95.28% Accuracy- 98.05% Specificity- 99.47%	Breast cancer records are available for 569 cases, of which 212 are malignant and 357 are benign.	Generative Adversarial Networks (GAN) and stacked autoencoders for disease prediction from EHRs.

IV. DISCUSSION

This paper provides a comprehensive review of current techniques that have been used in health prediction and monitoring using EMR data, with a focus on the integration of AI methodologies within EMR systems. The study highlights the potential of AI-driven approaches, including Deep Learning (DL), Machine Learning (ML), and Rule-Based Methods, in accurately diagnosing diseases. The paper discusses several instances where AI models have achieved predictive accuracies using the models in existing systems. According to the literature review of this paper, a few of the strengths and weaknesses were identified in each of these AI-driven approaches.

Because of their adaptability and capacity for probabilistic reasoning, machine learning techniques like support vector machines (SVM), Bayesian methods, and decision trees have been useful in the diagnosis of conditions like cancer, arthritis, and pulmonary hypertension. These techniques have also shown high predictive accuracies in a number of different disease types. SVMs are a powerful technology that are essential to data mining procedures since they support both linear and nonlinear regression techniques. SVMs are helpful for data prediction and classification, especially in the area of health research, because they can do binary and multiclass classification. SVMs do have significant drawbacks, too, such as the computationally demanding nature of model training and optimization. In contrast to more straightforward models like decision trees, they are also harder to interpret, which can be problematic in medical contexts when clarification is crucial. The Bayesian Network is another type of machine learning technology that has pros and limitations. Bayesian networks are helpful in controlling uncertainty and probabilistic reasoning in the setting of medical diagnostics, where ambiguity is common. They improve model predictions by combining prior information and experience. However, high-dimensional data is an issue for BNs, and as the number of variables increases, so does the model's complexity, making computation and interpretation challenging. Decision trees are machine learning techniques that give unambiguous decision paths and are highly interpretable. As a result, they can be used to justify diagnostic judgments in the healthcare industry. They perform effectively with numerical and categorical data, can adapt to various types of medical data, and can detect diseases early. However, decision trees are prone to overfitting, especially with complex or noisy data, and are limited in handling non-linear relationships compared to more sophisticated models like SVMs or deep learning techniques.

Rule-based approaches, which are noted for their speedy feature retrieval and simple knowledge representation, have exhibited excellent accuracy in specific disease diagnoses, such as colorectal cancer with an accuracy of 99.6% [22] and breast cancer with an accuracy of 90% [23]. However, they have some disadvantages, such as reliance on precise definitions, restricted flexibility and scalability, poorer accuracy in some circumstances, and difficulty understanding and implementing for non-experts. While rule-based systems have demonstrated excellent accuracy in certain areas and are praised for their simple logic, their rigidity and the requirement for precise data definitions can be significant limits, particularly in the dynamic and complicated field of healthcare.

Deep learning (DL) has demonstrated remarkable capabilities in biological applications. Because of its various processing levels, it is extremely successful at processing complex data, such as electronic medical records (EMR). DL approaches have shown great accuracy and sensitivity in a variety of medical activities, such as breast cancer detection, with an accuracy of 98.05% [33], although DL has certain limitations. Large datasets are often required for training, which can be a drawback in cases when data is sparse. Furthermore, training and implementing DL models can be computationally demanding, necessitating significant processing power and resources. Furthermore, DL models, particularly sophisticated structures, can lack interpretability, making it difficult to grasp the reasoning behind diagnoses or treatment decisions, which is critical in healthcare.

This review discusses different techniques for predicting cancer diseases, including SVM, Bayesian networks, rule-based methods, and stacked AE. Using SVM, Zhang et al. [9] classified cancer and achieved an accuracy of 97.33 %, while Zeng et al. [10] identified breast cancer with an accuracy of 93%. Using a Bayesian network, a similar cancer disease could be identified with 64.83% accuracy, while the same disease could be identified using Naive Bayes with 64.83% accuracy. Rule-based grammar was used to detect colorectal cancer [22], which earned an accuracy of 99.6%. Breast cancer was also detected using rule-based grammar [23], which achieved an accuracy of 90%. By combining AE and Softmax, Adem et al. of [32] classified cervical cancer with 97.25 percent accuracy. In this study, stacked AE and GAN [33] were used to predict breast cancer, and the accuracy rate was 98.05%.

To predict Asthma, different methods have been used. The Decision Tree and Fuzzy system [20] were used to diagnose and control asthma levels, and this system showed an accuracy of 90%. Wu et al. [28] used the RNN method to create a pediatric asthma prediction system with an accuracy f- a measure of 85.08%.

Even though there are many predictive models available, most of them are designed to predict single diseases without considering the many factors that can affect patients, for example, a cancer prediction system will only consider the symptoms of a patient to predict cancer and will not suggest other diseases based on these symptoms. However, several models have been developed to help identify multiple diseases, and this review discusses these systems. Al-Aidaroo et al. [16] classified and detected multiple diseases, involving hepatitis, cancer, and liver disorders, with an accuracy of 97.43%. With 86% and 84% sensitivity, [24] based on a rule-based algorithm, definite and probable Systemic lupus erythematosus (SLE) were detected. Shi et al [26] is another researcher focused on multiple diseases Cerebral infarction (CI), Pulmonary Infarction (PI), and Coronary Heart (CH) detected in this system, accuracy reached for each disease was CI 96.5%, PI 95.6%, CH 93.6%. Another system that is used to detect multiple diseases [21] is used to derive 78 diseases for this dataset taken from the Mount Sinai data warehouse of 7000 patients, this system received a 92.9% accuracy.

Data from the literature study indicates that certain approaches are more effective than others. While certain techniques may be more accurate for some illnesses but less accurate for others.

V. CONCLUSION

According to this review, several EMR system studies have been conducted recently to learn new facts about healthcare using technology. Using various procedures, EMRs provide a lasting record of patient care, reducing vulnerabilities and solving problems in modernized healthcare records.
Physicians can provide better care to patients when they have access to accurate and timely information. EMRs assist physicians in providing safer care, reducing medical errors, and improving the diagnosis of diseases. A competent EHR not only keeps track of patient allergies and medications but also checks for concerns when new medications are administered. An EMR can identify patterns of potentially related adverse outcomes and alert at-risk patients quickly. With the advancement of IT, EMR systems are now widely used to manage medical data and prescribe medication.
Different EMR systems using different techniques are installed and used in various healthcare facilities and these EMR systems have proven essential to delivering better patient care. In this review, it is classified into three primary categories machine learning, rule-based approach, and deep learning method which are then further subdivided depending on the suggested algorithm and have attempted to cover the most recent and current studies on autonomous diagnosis from electronic data. As discussed throughout the review, some methods can give accurate results in one type of disease, but not in another, and most systems are designed to predict and diagnose one specific disease, but very few systems have been able to detect multiple diseases simultaneously. According to the literature study, certain approaches were more effective than others.

Although EMR systems have their benefits, there are still some drawbacks, such as the need to update patient records after every appointment or consultation. Otherwise, physicians or clinical supervisors may later check the system and find incorrect information resulting in an inappropriate treatment plan. It is also possible that records may not be updated or inaccessible for an extended period if there is a power outage, location problems, or another issue. Another disadvantage is that they are still quite expensive.
Furthermore, future enhancements in EMR systems will include the ability to extract vital information from laboratory reports automatically. This integration of lab data with other EMR data will enrich the datasets used for predictions, leading to more accurate and comprehensive diagnostic insights. By encompassing a broader range of clinical information, including detailed lab results, these advanced systems will significantly refine the precision of disease prediction and patient treatment plans.

EHRs will be capable of handling massive amounts of data and complicated clinical test results in the future and eliminate current limitations and develop by using advanced existing methods and techniques to predict diseases more accurately. Related issues such as uncertainty in drawing conclusions and privacy issues will be addressed, and EHRs will come up with the genetic and behavioral data required for accurate prescribing and patient care improvement.

REFERENCES

[1] J. Wu, J. Roy, and W. F. Stewart, “Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches,” Med. Care, vol. 48, no. 6, pp. S106–S113, 2010.

[2] S. Ford, “Patient-centered Medicine, Transforming the Clinical Method,” Transforming the Clinical Method, vol. 7, pp. 181–182, 2004.

[3] M. A. Alkureishi, W. W. Lee, S. Webb, and V. Arora, “Integrating patient-centered electronic health record communication training into resident onboarding: Curriculum development and post-implementation survey among house staff,” JMIR Med. Educ, vol. 4, no. 1, 2018.

[4] J. Stausberg, D. Koch, J. Ingenerf, and M. Betzler, “‘Comparing paperbased with electronic patient records: Lessons learned during a study on diagnosis and procedure codes, ,’” J. Amer. Med. Inform. Assoc, vol. 10, no. 5, pp. 470–477, 2003.

[5] G. Makoul, R. H. Curry, and P. C. Tang, “‘The use of electronic medical records: Communication patterns in outpatient encounters,” J. Amer. Med. Inform. Assoc, vol. 8, no. 6, pp. 610–615, 2001.

[6] W. R. Hersh, “‘The electronic medical record: Promises and problems,” J. Amer. Soc. for Inf. Sci, vol. 46, no. 10, pp. 772–776, 1995.

[7] C.-S. Yu, Y.-J. Lin, C.-H. Lin, S.-Y. Lin, J. L. Wu, and S.-S. Chang, “‘Development of an online health care assessment for preventive medicine: A machine learning approach, ,’” J. Med. Internet Res, vol. 22, no. 6, 2020.

[8] Y. Si, “Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review,” J. Biomed. Inform, vol. 115, no. 103671, 2021.

[9] X. Zhang, J. Xiao, and F. Gu, “‘Applying support vector machine to electronic health records for cancer classification,” in Proc. Spring Simul. Conf. (SpringSim), 2019, pp. 1–9.

[10] Z. Zeng, “‘Contralateral breast cancer event detection using nature language processing,” in Proc. AMIA Annu. Symp, 2017.

[11] M. Jamaluddin and A. D. Wibawa, “Patient diagnosis classification based on electronic medical record using text mining and support vector machine,” in 2021 International Seminar on Application for Technology of Information and Communication, pp. 243–248.

[12] R. J. Carroll, A. E. Eyler, and J. C. Denny, “Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis,” AMIA Annu. Symp. Proc., vol. 2011, 2011.

[13] Y. Shen, “CBN: Constructing a clinical Bayesian network based on data from the electronic medical record,” J. Biomed. Inform, vol. 88, pp. 1–10, 2018.

[14] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, “‘Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system, ’’ BMC Med,” BMC Med. Informat. Decis. Making, vol. 6, no. 1, 2006.

[15] S. Sakai, K. Kobayashi, J. Nakamura, S. Toyabe, and K. Akazawa, “‘Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model, ’’ Methods Inf,” Methods Inf. Med, vol. 46, no. 06, pp. 723–726, 2007.

[16] K. M. Al-Aidaroo, A. A. Bakar, and Z. Othman, “Medical data classification with naive Bayes approach,” Inf. Technol. J.,

vol. 11, no. 9, pp. 1166–1174, 2012.

[17] J. Kazmierska and J. Malicki, "'Application of the Naïve Bayesian Classifier to optimize treatment decisions," Radiotherapy Oncol, vol. 86, no. 2, pp. 211–216, 2008.

[18] Y. Sun and D. Zhang, "'Diagnosis and analysis of diabetic retinopathy based on electronic health records," IEEE Access, vol. 7, pp. 86115–86120, 2019.

[19] A. Lungu, A. J. Swift, D. Capener, D. Kiely, R. Hose, and J. M. Wild, "'Diagnosis of pulmonary hypertension from magnetic resonance imaging-based computational models and decision tree analysis," Pulmonary Circulat, vol. 6, no. 2, pp. 181–190, 2016.

[20] A. Tyagi and P. Singh, "'Asthma diagnosis and level of control using decision tree and fuzzy system," Int. J. Biomed. Eng. Technol, vol. 16, no. 2, pp. 169–181, 2014.

[21] 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, 2016.

[22] H. Xu, "Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases," AMIA Annu. Symp. Proc, vol. 2011, pp. 1564–1572, 2011.

[23] C. Breischneider, S. Zillner, M. Hammon, P. Gass, and D. Sonntag, "Automatic extraction of breast cancer information from clinical reports,'' in Proc," IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS), pp. 213–218, 2017.

[24] A. Jorge, "'Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms," Seminars in Arthritis and Rheumatism, 2019.

[25] S. Mehrabi, "'Temporal pattern and association discovery of diagnosis codes using deep learning," in Proc. Int. Conf. Healthcare Informat, 2015, pp. 408–416.

[26] X. Shi, "'Multiple disease risk assessment with uniform model based on medical clinical notes," IEEE Access, vol. 4, pp. 7074–7083, 2016.

[27] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, Doctor AI: Predicting clinical events via recurrent neural networks. 2015.

[28] S. Wu, "'Modeling asynchronous event sequences with RNNs, ,'" J. Biomed. Informat, vol. 83, pp. 167–177, 2018.

[29] 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, pp. 1–10, 2016.

[30] L. Ali, C. Zhu, Z. Zhang, and Y. Liu, "Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network," IEEE J. Transl. Eng. Health Med., vol. 7, pp. 1–10, 2019.

[31] S. Farzi, S. Kianian, and I. Rastkhadive, "Diagnosis of attention deficit hyperactivity disorder using deep belief network based on greedy approach," in 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), 2017.

[32] K. Adem, S. Kilicarslan, and O. Cömert, "'Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder," Expert Syst. Appl, vol. 115, pp. 557–564, 2019.

[33] U. Hwang, S. Choi, H.-B. Lee, and S. Yoon, Adversarial training for disease prediction from electronic health records with missing data. 2017.

[34] J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review," IEEE Access, vol. 8, pp. 150489–150513, 2020.

ACKNOWLEDGMENT

I would like to extend my deepest gratitude to all the supervisors and lecturers who have significantly contributed to this research. Their guidance and support throughout this study have been invaluable.