A Comprehensive Review of Healthcare Prediction using Data Science with Deep Learning

Asha Latha Thandu, Pradeepini Gera
Department of Computer Science and Engineering, Koneru Laksmmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, 500302, India

Abstract—Data science in healthcare prediction technology can identify diseases and spot even the smallest changes in the patient’s health factors and prevent the diseases. Several factors make data science crucial to healthcare today the most important among them is the competitive demand for valuable information in the healthcare systems. The data science technology along with Deep Learning (DL) techniques creates medical records, disease diagnosis, and especially, real-time monitoring of patients. Each DL algorithm performs differently using different datasets. The impacts on different predictive results may be affects overall results. The variability of prognostic results is large in the clinical decision-making process. Consequently, it is necessary to understand the several DL algorithms required for handling big amount of data in healthcare sector. Therefore, this review paper highlights the basic DL algorithms used for prediction, classification and explains how they are used in the healthcare sector. The goal of this review is to provide a clear overview of data science technologies in healthcare solutions. The analysis determines that each DL algorithm have several negativities. The optimal method is necessary for critical healthcare prediction data. This review also offers several examples of data science and DL to diagnose upcoming trends on the healthcare system.

Keywords—Data science; deep belief network; healthcare; sparse auto encoder; deep learning

I. INTRODUCTION

Many studies have been conducted over the years on how to enhance the management and administration activities of the health sector and especially, healthcare offered to its patient [1]. Currently, the need of healthcare data is growing at an exponential rate in the healthcare system. From this point of view, the deployment of technology that is capable of being used in a creative way for the organization to help it achieve its goals is critical [2-5]. More preventive treatment options are becoming possible due to the use of health data analytics, especially predictive analytics. Despite access to a large amount of data, the healthcare industry lacks actionable knowledge that can be used to make predictions [6-10]. Despite its abundance, this is due to the fact that that health data is basically complex and fragmented [11]. Critical care, which is part of the health sector, faces the problem of increasing population and economic pressures, due to which it is difficult for most people to get the appropriate treatments. When talking to Intensive Care Unit (ICU) their condition often changes with every movement [12-14]. Likewise, with the advancement of technology in the healthcare sector an expectations of the patients are increasing but due to rising inflation, the required services cannot be provided. The main problem is to provide better, more effective care [15-17].

A data science solution for the analysis of healthcare data can help save patient lives and improve our quality of life [18-20]. Data science deals with several topics, such as data management and analysis, make correct decisions to improve the operation or system services (for instance, healthcare and transportation systems) [21-23]. In addition, with some very innovative and insightful techniques for displaying big data post-analysis, it is now easier to understand how any complex system works [24]. Due to the complexity of the healthcare system and operations, healthcare data are frequently fragmented. For example, different hospitals are only allowed to view clinical data of patients belonging to their specific patient groups [25]. These documents include highly Private Health Information (PHI) about specific individuals. This section examines the state-of-the-art in healthcare prediction using six deep representative architectures. Table I provides comparison of existing surveys with our survey.

The analysis shows that none of the studies carried out data science technology in healthcare prediction. Thus, this survey focuses on data science technologies in healthcare prediction. The contribution of this review is explained as follows:

- This review describes importance of data science technology in healthcare prediction in detail.
- Several deep learning technologies and applications used data science technology are analyzed for healthcare prediction.
- This survey helps researchers to diagnose the potential challenges in healthcare industry.

A. Data Collection

The searching keywords which are used in the above-mentioned databases like: “Healthcare Prediction” is demanded in every search abstract. It is a usual technique and consuming time as well. Also, searched for various synonyms and related keywords which meets the review outcomes such as, “Deep learning technology using data science technology”, “Applications in data science”, “Disease prediction using data science technology”, “healthcare monitoring” and “predictive analysis”. Again, refined the query to meet the particular outputs and again applied on Abstract and Title of the research article.
### TABLE I. COMPARISON OF EXISTING SURVEYS

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Description</th>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang Y et al. [77]</td>
<td>2018</td>
<td>Analyzed benefits of big data based on information technology infrastructure, managerial, organizational, operational, and strategic areas</td>
<td>This study employs a clear vision about how healthcare organizations are using big data analytics</td>
<td>The limitation of this study is source of data because of the IT adoption lags behind other industries in healthcare sector</td>
</tr>
<tr>
<td>Shirazi S et al. [73]</td>
<td>2019</td>
<td>Discussed data mining approaches and algorithms in healthcare domain</td>
<td>Classifying data mining papers regarding unsupervised and supervised learning algorithms</td>
<td>Didn’t give a clear vision of the applications of data mining technologies.</td>
</tr>
<tr>
<td>Bohr A and Memarzadeh K [78]</td>
<td>2020</td>
<td>Discussed major applications in artificial intelligence</td>
<td>Detailed research on applications directly related to healthcare and applications across the healthcare value chain including ambient assisted living and drug development</td>
<td>There is no detail about the different types of deep learning techniques</td>
</tr>
<tr>
<td>Li W et al. [79]</td>
<td>2021</td>
<td>Examined machine learning applications for big data in the healthcare sector</td>
<td>This survey examines basic big data concepts and relationship between big data and IoT</td>
<td>This review didn’t provide detail about disease prediction using ML techniques like covid-19, mental health</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>To offer an overview of data science technologies in healthcare prediction</td>
<td>Detailed review about various deep learning technologies and applications using data science technology</td>
<td>-</td>
</tr>
</tbody>
</table>

On searched articles, implements a quality assessment procedure following inclusion and exclusion constraints. Derived 3050 articles depends on the common keywords from several journals from different sources, such as Springer, ACM, IEEE Xplore, and Science Direct. Further, excluded some research papers that are not related to this study based on the heading. Fig. 1 shows flow chart of article selection procedure.

![Flowchart of article selection process](image)

**II. HEALTHCARE PREDICTION REVIEW METHODOLOGY USING VARIOUS DEEP LEARNING CLASSIFIERS**

In this section, healthcare prediction review methodology using deep learning classifiers are explained. Fig. 2 represents the taxonomy of the Healthcare prediction review.

**A. Healthcare Prediction using Data Science**

1) *Data acquisition:* HealthData.gov, Big Cities Health Inventory Data Platform, Chronic Disease Data, Human Mortality Database, Mental Disorders Datasets, MHealth Dataset, Medicare Provider Utilization and Payment Data, Life Science Database Archive, and WHO (World Health Organization) are some of the general datasets [26] used in this study to gather information on healthcare. The "precision medicine initiative" relates to healthcare. It sought to map the human genomes of one million United states residents, identify specific genetic defects that underlie a specific disease in a population, and effectively direct the development of drugs that are capable of precisely addressing a subcategory of molecular problems distributed by patients with a particular illness. Clinical, genetic features, and pathological can be included in IBM Watson for Healthcare, which can then provide standardized therapeutic paths and personalized therapy suggestions based on those features. By using DL algorithms, the enclitic firm (San Francisco, CA, USA) improves diagnostic performance in minimum time and in decreased costs using healthcare images (like MRIs and X-rays). Another excellent application is google flu trends, which uses monitoring data from laboratories around the US to forecast influenza-like disease than twice of doctor’s visits. Major research institutes, centers, and funding organizations have been able to invest in this field due to the significant role of these technologies in clinical and medical research. Along with health information systems, data science and DL techniques can be used to enhance healthcare management.
systems to meet the following objectives: cost reduction, fewer hospitalizations and shorter lengths of stays, prevention of fraudulent activity, classification in disease patterns, strong health insurance, and more efficient use of healthcare resources. The upcoming sections provide multiple application instances based on the various types of biomedical data, including biomedical time signals, biomedical images, and other biomedical information from wearable system, lab findings, and genomics. Modern biomedical equipment generates electrical signals from skin-mounted sensors, the features of which depend on the position of the sensor. These signals are an invaluable source of information for identifying and diagnosing diseases. By utilizing physiological signals like ElectroMyoGrams (EMG), ElectroEncephaloGrams (EEG), ElectroOculoGrams (EOG), and ElectroCardioGrams (ECG), DL can generate reliable applications.

In order to capture hierarchical relationship incorporated in deep features, DL has emerged as one of the primary study issues in the field of prognostics due to the fast growth of computing infrastructure. The deep network structure with numerous layers stacked in the network to completely capture the relevant data from the initial input data. In numerous domains, including image identification, audio recognition, and natural language processing, DL models have attracted significant interest and achieved notable successes. However, in the area of healthcare prediction, it has not yet been completely utilized. Six sample deep architectures—Deep Belief Network (DBN) [27], Convolutional Neural Network (CNN) [28], Recurrent Neural Network (RNN) [29], Long Short-Term Memory (LSTM) [30], Auto-encoder [31], and Sparse auto encoder [32]—were primarily the focus on the published research on DL. Based on these six exemplary deep architectures, this section aims to examine existing techniques.

The deep learning methods are, DBNs, CNN, RNN, LSTM, Auto encoders (AEs), and sparse Autoencoders (SAEs). Fig. 3 shows deep learning classifiers used in Health care prediction.

1) **Deep belief network**: A stack of Restricted Boltzmann Machines (RBMs) called the DBN consists of higher-order BMs as well as feature-identifying units on a single layer in BMs. The greedy layer-wise learning procedure of RBMs may pre-train the approach in an unsupervised manner with no limitations on the volume of training data.

A DBN is a generative statistical approach that can learns deep representations of input data. It stimulates the combined distribution of the hidden layers and the observable data. The RBM, an energy-based generative approach consisting of input layers, hidden layers, and symmetric networks between them, is one example of a basic, unsupervised network known as a DBN. In this situation, the incoming RBM uses the current hidden layer as its input layer.
The backpropagation technique may also be used to adjust the whole network. A rapid, layer-by-layer unsupervised architecture is created as a result of this composition, and it has since emerged as the most powerful DL algorithms. The capacity of DBN to recreate the input in an unsupervised manner is one of its standout features. For this reason, it has been used to carry out effective unsupervised feature learning in the fields of healthcare and transportation. For instance, the study used DBN to identify healthcare in input data. DBN is employed in this method to classify the health prediction data. DBN is the popular sequential modeling approaches, it has limits on long-term Remaining Useful Life (RUL) forecasts due to the inability of the trained network's weights to identify trends as a result of weight updates made when each input pattern was presented. Because of this, researchers have created the LSTM, which overcomes problems with long-term time dependence by employing input gates, forget gates, and output gates to manage information flow. In many sequential applications, the RNN and its variation, the LSTM networks, have gained considerable popularity. In recent years, healthcare prediction researchers have begun to investigate the potential of RNN, particularly LSTM. The extended kalman filter, truncated back propagation via time gradient computation, and evolutionary algorithms make up the RNN training algorithm. RNN are employed in this method to classify the health prediction data. An explanation of the classifier is given below.

There are various input layers (Y1, Y2), numerous hidden levels, and only one output layer in the recurrent neural network. The construction of RNN is seen in Fig. 5.

Healthcare detection is categorized using an RNN classifier. Healthcare input data are computed by RNN. There are several hidden layers to it. For the purpose of receiving output healthcare data, the network is connected to the proper
The input parameter for the network is weights and biases. The irrelevant data in each layer may be computed using the soft sign activation function as shown in Eq. (2).

$$f(y) = \frac{y}{1 + |y|}$$

(2)

In the above equation, input features of the network are denoted as y.

The hidden layer's output is given in Eq. (3),

$$l_s = g(Z_l y_s + V_{l_{s-1}} + a_l)$$

(3)

In the above equation, $Y_s$ represents input data, $l_s$ represents hidden layer, $Z_l$ and $V_l$ represents network weighted value, $a_l$ represents biases, $l_{s-1}$ denotes the previous state of the hidden layer and $g(\cdot)$ represents activation function.

By employing the Gaussian activation function, the determined outcome is given to the input of next layer to detect the healthcare data. Gaussian output is calculated as in Eq. (4).

$$Gaussian = e^{-y^2}$$

(4)

where, $Gaussian$ denotes the output of gaussian function, y be the variable. Then the total output is computed as in Eq. (5).

$$x_s = f(Z_l l_s + a_s)$$

(5)

In above equation, $x_s$ represents the network output, $l_s$ represents hidden layer, $Z_l$ represents network weighted value, $a_s$ represents biases and $f(\cdot)$ represents activation function. It results higher precision with lower execution time.

3) **Long short-term memory:** The LSTM network is a category of DL network created especially for linear data processing. An advantage of LSTM is their ability to retain both short-term and long-term values. Each cell in an LSTM unit has inputs, outputs, and forget gates. These three gates control the flow of information. By looking at the health prediction data, this LSTM categorizes the healthcare data in data science technology.

Numerous additional researchers have focused their attention on LSTM applications. LSTM, a kind of RNN for modelling long-term dependencies, was developed to solve the difficulties of time-series data. Like ordinary RNNs, LSTMs contain a memory for replicating the hidden layer activation patterns. Data processing involves hidden layer activations replicated iteratively.

The field of healthcare is paying more and more attention to LSTM models. According to sensor information such as blood pressure, temperature, and the results of lab tests, an LSTM model was employed as an example to diagnose healthcare data in a hospital setting. Similar to this, an LSTM model was applied to forecast examination outcomes based on prior measures. DeepCare is an LSTM-based system that is used to forecast future healthcare outcomes and infer the current sickness condition. A growing corpus of research has also focused on employing LSTMs to extract detailed data from medical texts like scientific publications, such as names of medications or medical occurrences.

4) **Auto-encoder:** The input data's encoder and decoder phases are rebuilt so that the Auto Encoder (AE) can learn a different representation of the information. As a result, it is frequently utilized for network pretraining. The stacked SAE, which combines multilayer AE such as denoising AE and SAE, is widely used Deep Neural Network (DNN) methods for handling the data. The main use of AE models in deep health monitoring is defect diagnosis. There are very few direct uses of AE in healthcare prediction in the literature; instead, AE is often employed to derive degradation features.

The unsupervised mode is applied when a hidden layer recreates the input layer in auto-encoders. Dimensions are allocated, in contrast to RNN, which incorporates weights and bias from the input layer to the hidden layer. The non-linear transformation function is utilized to calculate the stimulation of the hidden layer, which has lower dimensions compared to the input layer. Meanwhile, the input displays a dominating structure when the hidden layer size is decreased. To understand the identity function, non-linearity function should not be included, and the hidden layer and input layer dimensions should remain unchanged.

There are two subcategories of autoencoders. The first subcategory is denoising auto-encoder, and the approach is suggested to stop learning from becoming a simple problem. In this case, the input is rebuilt using noise that has been corrupted. Stacked auto-encoders are another form; they are created by stacking one auto-encoder layer on top of another. Each layer is trained separately to anticipate the output in healthcare applications, and the entire network is tweaked using supervised learning techniques. The architecture of AE is symmetrical, with the equal number of nodes in both the input layer and the output layer. The employment of AE has advantages such as dimensionality reduction and feature
learning. However, there are certain problems with dimension reduction and feature extraction in AE. The code of AE’s emphasis on minimizing data relationship loss results in the loss of several crucial data relationships. Its classification equation is given in Eq. (6).

$$\beta, \chi = \arg \min_{\beta, \chi} E(Y, (\chi \circ \beta)Y)$$ (6)

where, $\beta$ is represented as the encoder, $\chi$ is represented as the decoder, $E$ is represented as the error between input and output, $Y$ is represented as the input data, $\arg \min_{\beta, \chi}$ Minimum augmented of encoder and decoder.

5) **Sparse auto encoder**: The SAE is simply applying sparse constraints to the AE code. The unseen layer has more neurons than the input layer and output layer combined. While some of these neurons have a propensity to zero, others are similar to the input neurons. These are a few advantages of sparsity: The Sparse Auto Encoder’s Structure is depicted in Fig. 6.

- The model's anti-noise performance is enhanced over that of the original AE because it facilitates the extraction of key input characteristics.
- Sparseness is easier to understand and explain because it is satisfied by the majority of real-world circumstances.

The classification of healthcare prediction using SAE equation is given in Eq. (7).

$$SAE = E + \alpha \sum_i KL(\rho | \hat{\rho}_i)$$ (7)

where, $SAE$ represents the output of sparse autoencoder, $KL$ is represented as the kullback-leibler divergence, $KL$ divergence is known as the benchmark function that measures the effectiveness of different two disseminations are $\rho$ and $\hat{\rho}_i$. $\rho$ are represented as the anticipated neuronal network activity level, $\hat{\rho}_i$ indicates the average level of activation with $i^{th}$ neuron, $\alpha$ maintains the weight parameter.

C. **Comparison of Various Mechanisms on Healthcare Prediction**

In this section, comparison of different DL techniques is reviewed based on the strength and weakness of several DL models used for healthcare prediction. This review also considers accuracy term for comparison to determine the effectiveness of each model. Table II provides comparison of previously published research works advantages and their disadvantages.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Aim</th>
<th>Used technique</th>
<th>DL technique</th>
<th>Accuracy</th>
<th>Positives</th>
<th>Negatives</th>
<th>Scope for improvements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lu et al. [33]</td>
<td>Cardiovascular disease prediction</td>
<td>DBN</td>
<td>DBN</td>
<td>91.26%</td>
<td>Good stability</td>
<td>Minimum accuracy rate</td>
<td>Applying implemented model in terms of depth learning for cardiovascular prediction</td>
</tr>
<tr>
<td>Ali et al. [34]</td>
<td>Heart disease prediction</td>
<td>Optimally Configured and Improved DBN (OCI-DBN)</td>
<td>DBN</td>
<td>94.61%</td>
<td>Supports doctors to take efficient decisions</td>
<td>Not consider time complexity</td>
<td>Time complexity for the suggested technique will be computed because it most necessary factor in healthcare</td>
</tr>
<tr>
<td>Elkholy et al. [35]</td>
<td>Chronic kidney disease prediction</td>
<td>Modified DBN</td>
<td>DBN</td>
<td>98.5%</td>
<td>Beneficial for clinical decision making</td>
<td>Not support for unbalanced dataset</td>
<td>-</td>
</tr>
<tr>
<td>Javeed at al. [36]</td>
<td>Human behavior recognition model</td>
<td>Sustainable Healthcare Pattern Recognition (SPHR)</td>
<td>DBN</td>
<td>93.33% and 92.50%</td>
<td>Autonomous feature extraction reduces the dependency on domain expert</td>
<td>It gives lower accuracy for static activities</td>
<td>To perform the suggested model on complex activities</td>
</tr>
<tr>
<td>Pan et al. [37]</td>
<td>Heart disease prediction</td>
<td>Enhanced Deep learning assisted Convolutional Neural Network (EDCNN)</td>
<td>CNN</td>
<td>94.9%</td>
<td>Supports specialists to forecast information about heart patient using cloud platforms wherever in the world</td>
<td>Not applied in a real-world</td>
<td>Performance is further improved using feature selection techniques</td>
</tr>
<tr>
<td>Chung et al. [38]</td>
<td>Healthcare recommendations</td>
<td>CNN</td>
<td>CNN</td>
<td>90.1%</td>
<td>A dynamic cluster mechanism was suggested as well as prediction accuracy improved based</td>
<td>This recommendations system not suitable for symbolic</td>
<td>Further work can be extended to support symbolic knowledge expansion framework</td>
</tr>
<tr>
<td>Authors</td>
<td>Problem</td>
<td>Method</td>
<td>Features</td>
<td>Accuracy</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaur et al. [39]</td>
<td>Covid-19 detection</td>
<td>Deep CNN</td>
<td>CNN</td>
<td>92.93%</td>
<td>Appropriate for mobile applications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Younis et al. [40]</td>
<td>Brain tumor prediction</td>
<td>Visual Geometry Group (VGG16)</td>
<td>CNN</td>
<td>98.14%</td>
<td>The work uses MRI images to classify brain tumors and to support in making fast, effective and correct decisions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kim et al. [41]</td>
<td>Prediction of five chronic diseases</td>
<td>Character-RNN (Char-RNN)</td>
<td>RNN</td>
<td>77.6%, 82.6%, 80.6%, 82.5%, 96.5%</td>
<td>Efficiency for several chronic disease prediction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhu et al. [42]</td>
<td>Prediction of future glucose levels</td>
<td>Dilated Recurrent Neural Network (DRNN)</td>
<td>RNN</td>
<td>-</td>
<td>This approach performs better than existing forecasting algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feng et al. [43]</td>
<td>Healthcare prediction for football players</td>
<td>Smart football player health prediction approach</td>
<td>RNN</td>
<td>81%</td>
<td>Better reliability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ma et al. [44]</td>
<td>Parkinson’s chronic disease prediction</td>
<td>Self-attention and RNN</td>
<td>RNN</td>
<td>93.55%</td>
<td>Better Parkinson’s prediction skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carrillo-Moreno et al. [45]</td>
<td>Glucose predictor</td>
<td>LSTM</td>
<td>LSTM</td>
<td>-</td>
<td>To offer best prediction to patient, the suggested classifier validates several prediction times and input dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Said et al. [46]</td>
<td>Covid-19 detection</td>
<td>Bi-directional LSTM (Bi-LSTM)</td>
<td>LSTM</td>
<td>-</td>
<td>Better detection performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohamed et al. [47]</td>
<td>RNA mutation prediction</td>
<td>Seq2seq LSTM</td>
<td>LSTM</td>
<td>98.9% and 96.9%</td>
<td>The obtained outcomes illustrate the possibility of applying the LSTM network to RNA and DNA sequences in solving other sequencing problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algarni et al. [48]</td>
<td>Human emotion recognition</td>
<td>Stacked Bi-LSTM</td>
<td>LSTM</td>
<td>99.45%, 96.87% and 96.68%</td>
<td>The model’s performance results help to make precise medical decisions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mallick et al. [49]</td>
<td>Cancer detection using brain MRI images</td>
<td>Deep Wavelet Autoencoder (DWA)</td>
<td>AE</td>
<td>96%</td>
<td>Achieved good results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mahendran et al. [50]</td>
<td>Signal compression</td>
<td>priority-based convolutional AE</td>
<td>AE</td>
<td>-</td>
<td>Zero construction error</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mansour et al. [51]</td>
<td>Covid-19 prediction</td>
<td>Unsupervised DL based Variational AE</td>
<td>AE</td>
<td>98.7% and 99.2%</td>
<td>Good classification performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khamparia et al. [52]</td>
<td>Chronic kidney disease classification</td>
<td>Stacked AE</td>
<td>AE</td>
<td>100%</td>
<td>Good accuracy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- Several deep learning approaches and models will be implemented for further research work.
- Future work will be considered to differentiate brain tumor affected region from unaffected area precisely.
- Need to apply the suggested model to different ethnicities and lifestyle habits.
- In future, the suggested method will be embedded with IoT app.
- Adding few more parameters in the prediction system and computing the importance of these parameters for glucose prediction.
- Establishing lockdown relaxation scenario and analyze the impact of the relaxation.
- Applying new algorithms on several datasets to validate efficiency in emotion recognition.
- Integrating other variation of AE with DNN.
- Improving performance by focusing on the CNN framework.
- The suggested model embedded with IoT and cloud-based environment.
- Testing larger datasets for disease classification.
D. Multimedia Healthcare Applications using Deep Learning Technologies

An integration of several media or several types of data from multiple devices like texts, images, videos or audios called multimedia or multimodal data. To enhance the performance of the application, complementary information can be extracted from each modality by extracting multimodal data. Modality means encode information in a particular way. Various perspectives of a physiological objects using multimodal data provide additional information that can complement to the analysis. Table III provides multimedia healthcare applications using DL algorithms.

<table>
<thead>
<tr>
<th>Author</th>
<th>Applications</th>
<th>DL technique</th>
<th>Multimedia data</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alhussein M and Muhammad G [80]</td>
<td>Voice pathology prediction</td>
<td>Parallel CNN</td>
<td>Voice signals</td>
<td>Saarbrucken Voice Database</td>
</tr>
<tr>
<td>Algarni M et al. [48]</td>
<td>Emotion recognition</td>
<td>Bi-LSTM</td>
<td>EEG signal</td>
<td>DEAP dataset</td>
</tr>
<tr>
<td>Mukherjee D et al. [81]</td>
<td>Human activity recognition</td>
<td>EnsemConvNet</td>
<td>Time series of data</td>
<td>WISDM dataset, MobiAct dataset, UniMiB SHAR dataset</td>
</tr>
<tr>
<td>Yu Z et al. [82]</td>
<td>Disease Prediction</td>
<td>Deep Factorization Machine</td>
<td>Patients medical history</td>
<td>2020 artificial intelligence challenge preliminary competition</td>
</tr>
</tbody>
</table>

TABLE III. HEALTHCARE APPLICATIONS USING DL ALGORITHMS

E. Applications in Data Science Technology

Different applications of data science in healthcare prediction is explained in this section. This review considers different applications including speech therapy, disease detection, drug discovery, health monitoring, genomics and decision support system.

1) Speech therapy: Speech therapy supports children and adults affected with communication disorder to improve speech and languages. It supports with sound and voice production, early language skills, fluency and clarity. A speech therapist can use several types of therapy to support individuals with communication disorder related to fluency, speech, language, cognition, voice and swallowing. Mahmoud SS et al. (2020) [57] suggested assessment mechanism. Quadrature-based high-resolution time-frequency images with a CNN are used to detect the relationship between speech intensity and three speech intelligibility features in aphasic patients. The outcomes show a linear relationship with statistically significant correlations between the CNN model's normalized Truth-Class Output Functions (TCOA) and patients' pronunciation, tone scores, and fluency. Also, Bastanfard A et al. (2009) [58] proposed a new method that adopts image-based technique to combine visemes in persian by using coarticulation effect. The central frame was selected among various images for each phoneme defining various positions in different symbols. As a result of reconstruction, the weight value was established as criterion to compare viseme similarity. Experimental outcomes demonstrate the excellent precision and robustness of the suggested model.

2) Psychological prediction: This section explains specific DL algorithm used for disease detection including mental health, neurodevelopment disorder, and covid-19.

a) Mental health: In the real-word, the one of the most important and complex concern is auto diagnosis of mental health conditions. The mental health affects the way people behave, think and feel when they cooperate with world around them. Additionally, mental health issues are becoming a leading disability, contributing greatly to the global burden of disease. Du C et al. (2021) [59] designed a DL based Mental Health Monitoring System (DL-MHMS) for academy students. By using EEG signals, this suggested method used the effective CNN to categorize status of the mental health as normal, negative and positive. Zebertga K et al. (2023) [60] used Bidirectional Encoder Representations in Transformers (BERT) for excellently and efficiently recognizing anxiety and depression based posts by monitoring the context and semantic meaning of the words. Additionally, the knowledge distillation approach was proposed to transfer knowledge from a large pretrained model to a smaller model, which enhances accuracy. In last stage, BERT with Bi-LSTM and word2vec efficiently detects depression and anxiety symptoms.
b) Neurodevelopment disorders: Neurodevelopmental Disorders (NDs) affects brain functions as well as neurological developments, which causes problems in cognitive, social and emotional functioning. Some of the NDs are dyslexia, Autism Spectrum Disorder (ASD), and Attention Deficit Hyperactivity Disorder (ADHD). Sewani H et al. (2020) [61] offered an efficient prognosis of ASD using deep neural network particularly for children. This model integrates unsupervised learning, an AE and supervised DL using CNNs. The suggested approach performs better based on several validation and assessment measures. This model was tested on only 1112 rs-fMRI images. Minoofam SA et al. (2022) [62] suggested an adaptive reinforcement learning framework called RALF automatically generates content for students with dyslexia by Cellular Learning Automata (CLA). First, RALF creates samples of online alphabet as a simple form. The CLA system learns every instructions of character formation asynchronously using a reinforcement learning cycle. Then, generates persian words algorithmically. This stage determines the position of the character, cursiveness of the letters and the cell’s response to the environment. At last, RALF uses embedded word-generation approach to generate long-texts and sentences. Spaces between words are obtained using CLA neighboring states. The developed model offers many tests and games to enhance word pronunciation and people’s learning performance.

3) Covid-19 detection: More than five lakhs people died in India due to the covid-19. Generally, covid-19 often causes respiratory symptoms such as a cold, pneumonia or flu. Covid-19 can attack individuals’ lungs and respiratory system. Panwar H et al. (2020) [63] offered a DL algorithm for rapidly prediction of covid-19 cases depending on CT scan and X-ray images because the X-ray images offers necessary information in the covid-19 recognition. The model can identify covid-19 positive cases in less than two seconds, which is quicker than RT-PCR test. Kogilavani SV et al. (2022) [64] proposed several CNN designs like VGG-19, Densenet121, MobileNet, NASNet, Xception and EfficientNet to identify covid-19. This model didn’t detect covid-19 affected areas in the lungs.

4) Drug discovery: The importance of the drug discovery procedure is the advancement of novel drugs with potential interactions along with therapeutic targets. Drug discovery is depending on the conventional method, which focuses on holistic treatment. The medical communities of the world began to use the allopathic method to treatment and recovery in the last century. This change has led to victory in fighting against diseases, but has resulted in high healthcare burden and drug costs. Xiong Z et al. (2019) [65] introduced a novel graph neural network framework named Attentive FP to represent molecular, that used attention model to learn from suitable drug discovery datasets. The suggested approach achieves good performance on diverse datasets and that its learning is interpretable. The suggested Attentive FP automatically learns intramolecular interactions from specific tasks, helps derive chemical insights directly from data beyond human perception. ul Qamar MT et al. (2020) [66] analyzed the viral three-Chymotrypsin-Like cysteine protease enzymes, which is necessary for coronavirus lifecycle and controls its replication. This mechanism has been constructs 3D homology model by analyzing Chymotrypsin-Like cysteine protease sequences and screened it against a medicinal plant library comprising 32,297 potential anti-viral phytochemicals/ traditional Chinese medicinal compounds. The analysis demonstrates that the first nine hits the process of drug development to combat covid.

5) Healthcare monitoring: Data science serves important role in IoT. These devices consist of wearable devices that monitor heartbeat, temperature and medical parameters of the patients. Then, collected data is analyzed using data science. Based on the analytical results, doctors can able to monitor patient’s circadian cycle, their BP, and calorie intake. Ali F et al. (2021) [67] recommended BD analytics engine depending on data mining approaches, ontologies and Bi-LSTM. The data mining approaches are employed to preprocess the healthcare data and dimensionality reduction. The suggested ontologies are used to learn semantic knowledge about entities and features, and their relationships in Blood Pressure (BP) and diabetes domain. Finally, Bi-LSTM classifier is suggested to categorize effects from the drug side and abnormal states in individuals. The obtained results of the suggested method correctly handle big amount of data and enhances classification accuracy and prediction of drug side effect. M Abd El-Aziz R et al. (2022) [68] suggested IoT supported health monitoring system, which improves the data processing efficiency with the rapid adoption of cloud computing and expands data access in the cloud. The collected information from the real-time environment was stored in the cloud for data science processing. In this, improved pigeon optimization was suggested to combine the stored data in the cloud that supported for enhancing prediction rate. Following that, feature extraction and selection are performed with the help of optimal feature selection approaches. To categorize human healthcare, Backtracking Search-Based DNN (BS-DNN) was suggested.

6) Genomics: Genomics is the study of the sequencing and analysis of genes. A genome contains DNA and all the genes of an organism. Since the compilation of the Human Genome Project, research has progressed rapidly and embedded itself in the data science and big data fields. Nasir MU et al. (2022) [69] suggested CNN based AlexNet to predict genome multi-class disorder for developing Advance Genome Disorder Prediction Model (AGDPM). This model was capable of genome disorder prediction and it uses large amount of data to processes patient’s genome disorder data. The suggested prediction system improves biomedical system based on predict genetic disorders and reduces high mortality rates.

7) Decision support system: The goal of a Decision Support System (DSS) is to enhance healthcare delivery through improving clinical decisions with targeted clinical knowledge, patient data and other health related information.
Malmir B et al. (2017) [70] presented a DSS called Fuzzy Expert System (FES) to support doctors for better decision making in medical diagnosis. This suggested model conducts a cross-sectional study to gather information about diseases by asking clinicians on all signs based on diseases. According to this information, then fuzzy rule-based system with the necessary symptoms necessary signs based on the suspected disease was developed. To prove effectiveness of the suggested method, two case studies conducted on kidney stone and kidney infection. Khiabani SJ et al. (2022) [71] developed DSS in terms of neural network and statistical process control charts for identifying and control of Myocardial Infarction (MI) and continuous observing of patient BP. A group of patients was used to prove the suggested system’s efficiency. The outcomes validate that the suggested model can detect MI by the parameter of accuracy and precision.

III. RESULTS AND DISCUSSIONS

A review of data science and deep learning techniques for healthcare prediction with high accuracy is discussed in this section.

Fig. 7 illustrates the various deep learning techniques contribute to covid-19 prediction by the parameter of accuracy. The analysis demonstrate that the AE performs over other deep learning techniques including SAE, and CNN. But it is important to note that each work uses different datasets to prove its effectiveness. Correspondingly, the work based on AE [51] predicts covid-19 using chest x-ray images. It is necessary to diagnose covid-19 using other modalities such as CT, MRI and so on.

![Accuracy Chart](image)

**Fig. 7.** Comparison of covid-19 detection using reviewed techniques.

Table IV shows the comparison of DL techniques in terms of precision and sensitivity for heart disease prediction. The analysis shows that the EDCCNN technique occurs better precision than other techniques. For the sensitivity analysis, the authors [54] uses optimized feature learning techniques, this supports to achieves a highest sensitivity of 100%.

Even though AI-based DL are efficiently predicting the data, some of the classifiers reduce the accuracy due to their limited data size, high dimensionality, efficient feature selection technique, model generalizability, and clinical implementation. The explanations for these limitations are given as follows:

1) **Limited data size:** One of the challenges facing this study was inadequate data for training the classifier. The less input data size allows a number of the training set, which reduces the accuracy of the presented methods. Hence, to improve the accuracy of the training samples and to train a large number of datasets new methods are used, which performs better than previous classifiers [72].

2) **High dimensionality:** High dimensionality is another problem faced by the previous methods. The input data set consists of number of data with a high number of features. The current methods face several high dimensionality issues for extracting the features. Hence, new feature-extracting methods are used to overcome these issues [73].

3) **Efficient feature selection technique:** Previously several feature selection and disease prediction methods were employed for predicting the disease in its early stage. But they are limited due to high computational complexity. To overcome this limitation, the most effective feature selection methods and pre-processing methods are used for predicting the data’s higher accuracy [74].

4) **Model generalizability:** For improving the prediction results, a change in the research is needed based on the model’s generalizability. Generalizability is used to analyze the results from highly populated situations with prediction methods. Previously there were several prediction methods for predicting the patient data in a single site. Nowadays, it is required to predict the patient data in multiple sites, and while improving the predicting data in multiple sites the model’s generalizability is enhanced [75].

5) **Clinical implementation:** Finally, AI-based ML and DL methods have provided good results for predicting disease in DS in healthcare predictions. But still, this method faces issues in practical implementations with clinics that are not supported. In the current work, these AI methods are required to validate the clinical setting for assisting the doctor in confirming the findings and decisions [76].

**TABLE IV. COMPARISON OF DIFFERENT PERFORMANCE METRICS BASED ON HEART DISEASE PREDICTION**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Precision</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBN [33]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OCI-DBN [34]</td>
<td>93.55%</td>
<td>96.03%</td>
</tr>
<tr>
<td>EDCNN [37]</td>
<td>99%</td>
<td>97.51%</td>
</tr>
<tr>
<td>Stacked SAE [54]</td>
<td>94.38%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Many clinical elements are involved in EHR-based systems. There are millions of data points available for this. It would not be easy to manage and regulate the whole data of millions of people. There are several critical challenges yet to be overcome:

- There was a lot of unorganized or inaccurate data, making it tough to gain a more profound knowledge of it.
• It is difficult to strike the right balance between the preservation of patient-centric data and the excellence and convenience of this information.

• Keeping data private, storing it efficiently, and transferring it requires a lot of workforces to ensure that these requirements are met continuously.

• Lack of language proficiency when handling data.

IV. CONCLUSION

This manuscript proposes a review of data science and healthcare prediction in data science technology in order to forecast healthcare with high accuracy is successfully predicted. In this, the healthcare prediction is classified using deep learning classifiers based on health issues such as lab reports, medical imaging and EHR. In this, the deep learning classifiers are CNNs, RNN, LSTM, RBMs, DBNs, AEAs and SAE. In healthcare services, the analysis shows that the existing approaches are not efficient to handle big data. Existing and future smart healthcare systems requires examinations about the design considerations such as maintainability, performance, accuracy, scalability, cost, security, responsiveness, fault tolerance, and reliability. It’s hoped that this review paper will help many scholars to improve their knowledge for their future research work. Future research is planned to review machine and deep learning techniques with an approach that optimises healthcare data in different environment like IoT, cloud and so on.

REFERENCES


