PERSONALIZED MEDICINE RECOMMENDATION FOR DIABETIES USING DATA MINING TECHNIQUES

Dr.S.Kalarani,
Professor
Department of Information Technology
St Joseph's Institute of Technology
OMR, Chennai, India

Aakash S K
B.Tech IT-Final Year
Department of Information Technology
St Joseph's Institute of Technology
OMR, Chennai, India

***Abstract* -Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. With the recent advancements in analyzing high volume, complex and unstructured data, modern learning methods are playing an increasingly critical role in the field of personalized medicine. Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore deep leaning algorithm along with three machine learning classification algorithms Decision Tree, SVM and ordinal logistic regression are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show ordinal logistic regression outperforms with the highest accuracy of 76.30% comparatively other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner. After the analysis we have a research in the field of IoT based wearable which automatically releases required amount of insulin. The day to day insulin requirement is calculated based on food intake and physical activities done by the patient. Based on this exact amount of insulin dosage will be injected into the body.**

Keywords—Personalized medicine, data mining

1. INTRODUCTION

Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both.

Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D).

Patients with type 1 diabetes are normally younger, mostly less than 30 years old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels. This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy. Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases.

Diabetes is one of the most common human diseases and has become a significant public health concern worldwide. There were approximately 450 million people diagnosed with diabetes that resulted in around 1.37 million deaths globally in 2017. Diabetes patients are at elevated risk of developing health complications such as kidney failure, vision loss, heart disease, stroke, premature death, and amputation of feet or legs, which can lead to dysfunction and chronic damage of tissue. In addition, there are substantial economic costs associated with the disease. Additionally, there could be productivity loss due to diabetic patients in the workforce. An individual at high risk of diabetes may not be aware of the risk factors associated with it. Given the high prevalence and severity of diabetes, researchers are interested in finding the most common risk factors of diabetes, as it could be due to a combination of several reasons. Determining the risk factors and early prediction of diabetes have been vital in reducing diabetes complications and economic burden and is beneficial from both clinical practice and public health perspectives. Similarly, studies find that screening high-risk individuals identifies the population groups in which implementing

Decision tree is a basic classification and regression method. Decision tree model has a tree structure, which can describe the process of classification instances based on features. It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature space and class space. Decision tree uses tree structure and the tree begins with a single node representing the training samples. If the samples are all in the same class, the node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree. According to the value of the current decision node attribute, the training samples are divided into several subsets, each of which forms a branch, and there are several values that form many branches. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples.

The typical algorithms of decision tree are ID3, C4.5, CART and so on. In this study, we used the J48 decision tree in WEKA. J48 another name is C4.8, which is an upgrade of C4.5. J48 is a top-down, recursive divide and conquer strategy. This method selects an attribute to be root node, generates a branch for each possible attribute value, divides the instance into multiple subsets, and each subset corresponds to a branch of the root node, and then repeats the process recursively on each branch. When all instances have the same classification, the algorithm stop. In J48, the nodes are decided by information gain. According to the following formulas, in each iteration, J48 calculates the information gain of each attribute, and selects the attribute with the largest value of information gain as the node of this iteration.

Attribute A information gain:

$$Gain(A)=Info(D)-InfoA(D)Gain(A)=Info(D)-InfoA(D)..(1)$$

Pre-segmentation information entropy:

$$Info(D)=Entropy(D)=-\Sigma jp(j|D)logp(j|\ |\ |\ d)Info(D)=Entropy(D)=-\Sigma jp(j|D)log\,p(j|d)\ldots. \quad (2)$$

Distributed information entropy:

$$InfoA(D)=\Sigma i=1vninInfo(Di)InfoA(D)=\Sigma i=1vninInfo(Di)..(3)$$

Data mining is a field of research that has emerged in the 1990s, and is very popular today, sometimes under different names such as “big data” and “data science “, which have a similar meaning. To give a short definition of data mining, it can be defined as a set of techniques for automatically analyzing data to discover interesting knowledge or pasterns in the data.

The reasons why data mining has become popular is that storing data electronically has become very cheap and that transferring data can now be done very quickly through the fast computer networks that we have today. Thus, many organizations now have huge amounts of data stored in databases, that needs to be analyzed.

Having a lot of data in databases is great. However, to really benefit from this data, it is necessary to analyze the data to understand it. Having data that we cannot understand or draw meaningful conclusions from it is useless. So how to analyze the data stored in large databases? Traditionally, data has been analyzed by hand to discover interesting knowledge. However, this is time-consuming, prone to error, doing this may miss some important information, and it is just not realistic to do this on large databases.

We consider a combined approach of ordinal logistic regression and machine learning to predict the risk factors of type 2 diabetes mellitus. The ordinal logistic regression compares several prediction models for predicting diabetes.

The rest of the paper is planned as follows: the second section provides an overview of the proposed system. The section that follows presents results and analyses. Then, we show the comparison of the state-of-the-art techniques. At last, Section 5 concludes the paper.

II. Proposed Methodology

Personalized medicine, alternatively called precision medicine, is the tailored medical treatment to different groups of patients based on their predicted response to a risk or disease. As Hippocrates stated, “It is far more important to know what sort of person the disease has than what sort of disease the person has”. This catches the essence of personalized medicine, where the focus is on the patient- his genetics, inheritance, lifestyle, etc., instead of on the disease phenotypes, which treats an “average patient” with a “one-size-fits-all” medical procedure. With the growing size of medical database characterizing the patient response and genetics, personalized medicine is becoming an increasingly viable alternative to traditional medication based on – omics data. Personalized medicine now has wide applications in the treatment of complex diseases that depends highly on genetics, environment and medical history. Such diseases include diabetes, cancer, heart disease and psychiatric disease. One challenge in personalized medicine is how to predict the clinical response of prescribed drugs to each individual patient with a highly-diversified problem. For example, in the treatment of cardiovascular disease, it has been known that patient response varies dramatically.

The adoption of newest developments in computer algorithms can be promising, by providing tools to discover data patterns that can be otherwise hidden and difficult to observe. Such findings of new patterns may have deep implications for realizing the biological origin and the underlying pathways of a disease for a specific subgroup of people. Moreover, personalized medicine entails a large amount of data, from age, weight, blood pressure, to medical history and genomic data, all different for each patient. To deal with such an amount of data in order to generate a medical treatment plan often exceeds what a doctor can do based on experience. The data are not only large in size and dimensionality, but also unstructured and inhomogeneous.

Several factors might be related to diabetes, including blood pressure, pregnancies for women, age and body mass index, etc. As a component of diabetes management, it would be helpful to know which variables are related to diabetes.

In this paper, we propose a diabetes prediction system for better diagnosis. Our work focuses on the following points: (1) Set up a system architecture for diabetes prediction based on DNN algorithm in order to make an efficient decision to the diabetes diagnosing; • an evaluation of four different DNN architectures to get the best model.
(2) A comparison of best DNN model’s results against those of many well-known ML classifiers such as LR, SVM, XGBoost, DT, and RF.
(3) Furthermore, we compare our proposed method with the state-of-the-art methods that used the same datasets, the same experimental protocol, and the same performance measurements.

III. RELATED WORK

Biomarker development in the precision medicine era: lung cancer as a case study. Vargas AJ, Harris C,2016 Precision medicine relies on validated biomarkers with which to better classify patients by their probable disease risk, prognosis and/or response to treatment. Although affordable 'omics'-based technology has enabled faster identification of putative biomarkers, the validation of biomarkers is still stymied by low statistical power and poor reproducibility of results.

This Review paper summarizes the successes and challenges of using different types of molecule as biomarkers, using lung cancer as a key illustrative example. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Kim J,Shin H 2013 . Prognostic studies of breast cancer survivability have been aided by machine learning algorithms, which can predict the survival of a particular patient based on historical patient data. However, it is not easy to collect labeled patient records. It takes at least 5 years to label a patient record as 'survived' or 'not survived'. Unguided trials of numerous types of oncology therapies are also very expensive. Confidentiality agreements with doctors and patients are also required to obtain labeled patient records.

IV System Architecture

Existing System- In the existing system, the doctors determine the blood sugar level or glycosylated hemoglobin level by undergoing a Fasting Blood Sugar(FBS) test or AC1 test. According to the results they instantly suggest medicines for the patients.

Disadvantages-Drug adverse reactions for patients cannot be determined accurately. Drug recommendations are common for all patients irrespective of their working nature and life style. Patients historical records are hard to maintain and manipulate.

4.1 Advantages:

Reduction of mortality rate,Improvement prognosis , Easy to identify disease progression at the early stages.With the growing size of medical database characterizing the patient response and genetics, personalized medicine is becoming an increasingly viable alternative to traditional medication. Personalized medicine now has wide applications in the treatment of complex diseases that depends highly on genetics, environment and medical history. Such diseases include diabetes, cancer, heart disease and psychiatric disease.

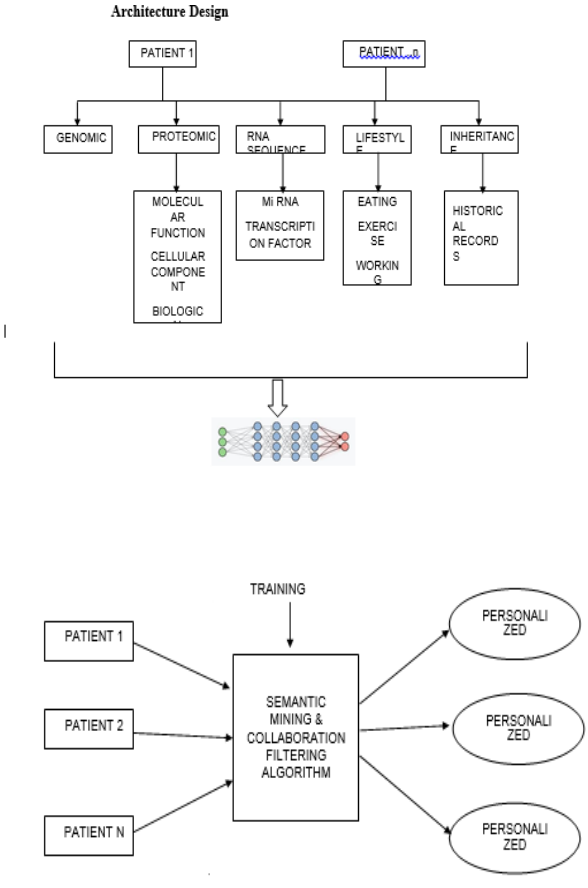
Our proposed architecture, we are working with omics data such as – genomic, proteomic and metabolomics through which we recommend a tailored medical treatment by performing an effective analysis of DNA, RNA sequences and lifestyle.

4.2 Description

The architecture diagram shown above represents the personalized medicine recommendation system that entails a large amount of data such as age, weight, blood pressure, medical records, genomic, proteomic, RNA sequences and lifestyle which are all different for each patient according to their age, working nature and circumstances. The proteomic data includes a collection of molecular function, cellular component and biological process for performing cross ontology process. Using these the Mi RNA sequence could be identified which would be further helpful for prescribing lifestyle for patients. All the data that are fetched are provided as input to the training model, which effectively demonstrates high potential to relate the medical history, laboratory data, genotype data, and familial inheritance data of biomedical research. After dealing with large amount of data, the training model suggests hidden relationship with these data the patter matching is performed by sematic mining for knowledge discovery when used in combination with proteomics, genomics and metabolomics data. This can make better informed medical decisions by uncovering the expressed molecular activities. Apart from this the drug adverse reactions and its sensitivity could be determined. This leads to classification of patients into different subgroups.

Testing of drugs is based on,Testing drug sensitivity based on patient genetics profile, especially in diabetes treatment. Predicting drug indications based on drug-drug, disease-disease and drug- disease similarity

Analyzing drug adverse side reactions.The information about these drug responses are also stored for future use.



SYSTEM IMPLEMENTATION

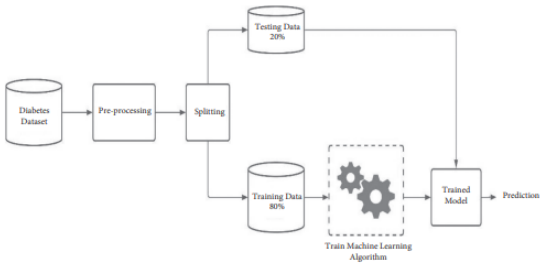


FIGURE 1: The proposed system flowchart.

4.3 Ordinal Logistic Regression

Ordinal Logistic Regression. Logistic Regression (LR) is a subset of generalized linear models which deals with the analysis of binary data, which seeks out the best-fitting model for describing the connection between dependent and independent predictors. When it comes to predicting sickness or health status, the LR model is most commonly used. Based on the risk factors given, the LR model can calculate the likelihood of an individual acquiring diabetes disease. If a person suffers from diabetes disease, the value of target is 1; otherwise, target is 0. We determined that the probability of an individual developing diabetes disease is P (X).

The LR model's formula is defined as follows:

$$\text{logistic}(p) = \ln\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k. \quad (2)$$

After exponentiating both sides, we obtain

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}. \quad (3)$$

The probability of an individual developing diabetes disease can be written as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}, \quad (4)$$

where $X_1, X_2, X_3, \ldots, X_k$ represent the risk factors and $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are regression coefficients.

Support Vector Machine. SVM is a nonprobability classifier with a separating hyper plane as its formal definition. *e technique creates an ideal hyper plane with the greatest distance from the support vectors based on the available training data (supervised learning). This hyper plane is a line that divides a plane into two classes in two-dimensional space. The epsilon ε, regularization, and kernel parameters are the SVM classifier's tuning parameters. The principle of SVM is shown below in Figure 2.

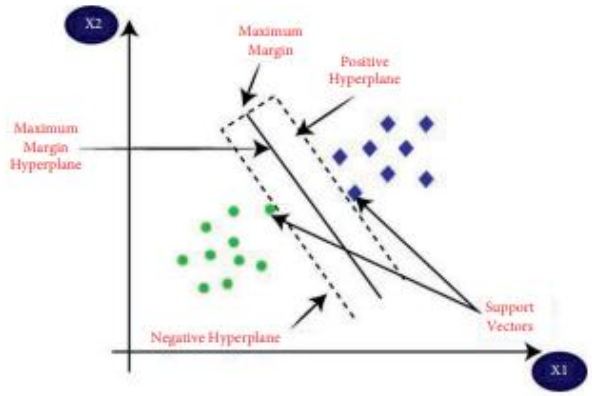


FIGURE 2: Support Vector Machines (SVM) [6].

4.4 Experimental Results

In this section, we evaluate the performance of DNN algorithm by using the testing data to assess the effectiveness of our system based on several evaluation metrics. Besides, comparison between our proposed model and the machine learning algorithms described, has been conducted in order to demonstrate the superiority of our model.

The used dataset was split into two subsets, the first one for training which contains 80% of the whole data (547 diabetics/1053 no diabetics) and the other for testing which contains 20% of the whole data (137 diabetics/263 no diabetics).

4.5 Evaluation Metrics.

The confusion matrix is considered as a great tool to show the results summary of a model with the classification issues. In the classification, the prediction can be one of four special cases as follows. If the actual value of the target in the dataset is True and the classifier predicts it as such, then the prediction is a True Positive (TP). On the contrary, if the classifier predicts it as False, then the prediction is a False Negative (FN). Similarly, if the actual value of the target in the dataset is False and the classifier predicts it as such, then the prediction is True Negative (TN). On the contrary, if the classifier predicts it as True, then the prediction is False Positive (FP). Finding out how the developed predictive model performs becomes easy with the help of the confusion matrix, which is clearly shown in Fig

		predicted values	
		0	1
Actual values	0	True Negative (*TN*)	False Positive (FP)
	1	False Negative (FN)	True Positive (*TP*)

FIGURE 6: Confusion matrix for binary classification.

The following metrics are used to evaluate the proposed model.

Accuracy (Acc) : is the percentage of the correct predictions that a classifier has made compared with the actual values of the target in the testing phase.

Acc = [(TP + TN) / (TP + TN + FP + FN)] * 100

Sensitivity (Sens) : gives information about the percentage of True Positives that are correctly classified during the test.

Sens = TP (TP + FN) ∗ 100%.

Specificity (Spec) : gives information about of True Negatives that are correctly classified during the test.

Spec = TN(TN = FP) * 100

Precision (Pre): is the percentage of instances that a classifier has labelled as positive with respect to the total predictive positives (the exactness of a classifier).

Pre = TP (TP + FP) ∗ 100%.

F1-score shows the harmonic mean of precision and recall.

F1 − score= 2 ∗ TP (2 ∗ TP + FN + FP) ∗ 100%.

Prediction with ML Methods. A comparative analysis of all the conventional machine learning algorithms has be done in this section for diabetes prediction. It has been done for comparing and analyzing accuracies of all the conventional algorithms.

4.6 Total Daily Insulin Requirement

The general calculation for the body's daily insulin requirement is:

Total Daily Insulin Requirement (in units of insulin)
= Weight in Pounds ÷ 4

Alternatively, if we measure your body weight in kilograms:

Total Daily Insulin Requirement (in units of insulin)
= 0.55 * Total Weight in Kilograms

Example 1:

If we are measuring your body weight in pounds:

Assume you weigh 160 lbs.

In this example:

TOTAL DAILY INSULIN DOSE = 160 lb ÷ 4 = 40 units of insulin/day

Example 2:

If we are measuring your body weight in kilograms:

Assume your weight is 70Kg

In this example:

TOTAL DAILY INSULIN DOSE = 0.55 x 70 Kg = 38.5 units of insulin/day.

If our body is very resistant to insulin, you may require a higher dose. If our body is sensitive to insulin, we may require a lower insulin dose.

Basal/Background and Bolus Insulin Doses. Next, we need to establish the basal/background dose, carbohydrate coverage dose (insulin to carbohydrate ratio) and high blood sugar correction dose (correction factor).

Basal/background insulin dose:

Basal/background Insulin Dose = 40-50% of Total Daily Insulin Dose

Example

1. Assume you weigh 160 pounds
2. Our total daily insulin dose (TDI) = 160 lbs ÷ 4 = 40 units.

In this example:

Basal/background insulin dose = 50% of TDI (40 units) = 20 units of either long acting insulin, (such as glargine or detemir) or rapid acting insulin if you are using an insulin pump (continuous subcutaneous insulin infusion device).

The carbohydrate coverage ratio:

500 ÷ Total Daily Insulin Dose = 1 unit insulin covers so many grams of carbohydrate

This can be calculated using the Rule of "500": Carbohydrate Bolus Calculation

Example:

Assume your total daily insulin dose
(TDI) = 160 lbs ÷ 4 = 40 units

In this example:

Carbohydrate coverage ratio = 500 ÷ TDI (40 units) = 1unit insulin/ 12 g CHO

This example above assumes that we have a constant response to insulin throughout the day. In reality, individual insulin sensitivity varies. Someone who is resistant in the morning, but sensitive at mid-day, will need to adjust the insulin-to-carbohydrate ratio at different meal times. In such a case, the background insulin dose would still be approximately 20 units; however, the breakfast insulin-to-carbohydrate ratio might be breakfast 1:8 grams, lunch 1:15 grams and dinner 1:12 grams.

4.7 Glucose Level Monitoring

Eversense is a subcutaneous implant that continuously monitors blood glucose levels. Eversense -Developed by the US company Senseonics. It initially needs to be installed under the skin by a doctor, the sensor can work for up to three months before needing a replacement. Eversense measures glucose in the interstitial fluid under the skin of the upper arm by using a polymer that fluoresces in response to the levels of blood sugar.

The data is then sent to a transmitter that displays the blood glucose levels in real time.

It simply a sensor which continuously monitor human sugar level and send a remainder for insulin dosage through smart phone. But many research going on for wearable with loaded insulin device. The Eversense implantable sensor senses the glucose level continuously and record the data. A deep learning based AI system evaluates the exact dosage requirement of the human being based on carbohydrate and other intakes along with physical activities.

This system works with smart phone app and it give command to the wearable band , to release exact dosage of insulin to the body. After receiving the alert just tighten the band and start initiate the insulin delivery. A small needle

ejected from the device and pricks the flush part of the body and releases the exact dosage as prescribed by the mobile app.

V CONCLUSION

Identifying individuals at high risk of developing diabetes is a critical component of disease prevention and healthcare. This study presents a predictive equation of diabetes to provide a better understanding of risk factors that could assist in classifying high-risk individuals, make the diagnosis, and prevent and manage diabetes. Five critical variables identified in predicting type 2 diabetes are age, BMI, pedigree, glucose and frequency of pregnancies. We conclude that our proposed model has a prediction accuracy of 78.26%, with a cross-validation error rate of 22.86%. As for the case of a classification tree, we would choose the tree with six nodes since it has the highest prediction accuracy (74.48%) than other possible sub trees. The results imply that if we control these five predictors by taking the necessary steps, it could lower type 2 diabetes prevalence. In addition, accurately

Predicting diabetes might help design interventions and implement health policies that may aid in disease prevention. Suggestions are recommended to calculate exact requirement of insulin as the intake and physical activities. A mobile application with deep learning algorithms are developed as an alert system. An IoT implantable sensor with insulin loaded devices are planning to release exact amount of insulin into the patient as wearable device.

The development of personalized medicine in areas such as drug development, disease characteristics identification, and therapeutics effect prediction would have a greater impact in human life's and health care sector providing advantages such as identification of disease at early stage, reduces mortality rate and improving prognosis

The topic of personalized medicine has been attracting numerous attention from many researchers over the past decades. Thus in future work considering disease characteristics including research on susceptibility, occurrence, recurrence, survivability and phenotypes during drug recommendation will make medical decision making accurate and easy.

REFERENCE

[1] A. J. Vargas and C. C. Harris, "Biomarker development in the precision medicine era: lung cancer as a case study," Nature Reviews Cancer, vol. 16, no. 8, p. 525, 2016.

[2] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," Journal of the American Medical Informatics Association, vol. 20, no. 4, pp. 613–618, 2013.

[3] E.M.AntmanandJ. Loscalzo, "Precision medicine in cardiology," Nature Reviews Cardiology, vol. 13, no. 10, p. 591, 2016.

[4] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: Opportunities and challenges," Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2017.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.

[6] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," Molecular pharmaceutics, vol. 13, no. 5, pp. 1445–1454, 2016.

[7] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," Molecular informatics, vol. 35, no. 1, pp. 3–14, 2016.

[8] J. Davis, E. Lantz, D. Page, J. Struyf, P. Peissig, H. Vidaillet, and M. Caldwell, "Machine learning for personalized medicine: Will this drug give me a heart attack," in the Proceedings of International Conference on Machine Learning (ICML), 2008.

[9] P. Kieseberg, H. Hobel, S. Schrittwieser, E. Weippl, and A. Holzinger, "Protecting anonymity in data- driven biomedical science," in Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer, 2014, pp. 301–316.

[10] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 129–136.

[11]V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," Jama, vol. 316, no. 22, pp. 2402–2410, 2016.

12]J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[13]A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 215–223.

[14]F. Movahedi, J. L. Coyle, and E. Sejdic, "Deep belief networks´ for electroencephalography: A review of recent contributions and future outlooks," IEEE journal of biomedical and health informatics, 2017.

[15]H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," Drug discovery today, 2018.

[16]S. Ekins, "The next era: Deep learning in pharmaceutical research," Pharmaceutical research, vol. 33, no. 11, pp. 2594–2603, 2016.

[17]M. H. Segler and M. P. Waller, "Neural-symbolic machine learning for retro synthesis and reaction prediction," Chemistry-A European Journal, vol. 23, no. 25, pp. 5966–5971, 2017.

[18]T. Jo, J. Hou, J. Eickholt, and J. Cheng, "Improving protein fold recognition by deep learning networks," Scientific reports, vol. 5, p. 17573, 2015.

[19]L. Deng and Y. Liu, Deep Learning in Natural Language Processing. Springer Singapore, 2017.

[20]Y. LeCun et al., "Lenet-5, convolutional neural networks," URL: http://yann. lecun. com/exdb/lenet, p. 20, 2015.

[21]J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo- labeled patient data," Journal of the American Medical Informatics Association, vol. 20, no. 4, pp. 613– 618, 2013.