Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 01 January 2024

An Experimental Study of Protein Classification in Artificial Neural Networks for Cancer Diagnosis

Anil Kumar Sharma, Prof. (Dr.) Pushpneel Verma
Research Scholar, CSE Dept., Bhagwant University, Ajmer, Rajasthan
Professor, CSE Dept., Bhagwant University, Ajmer, Rajasthan

***Abstract*:** To study protein classification in Artificial Neural Network, we must understand some important concepts. The purpose of protein classification is to classify proteins on the basis of their form and function. It is particularly useful in bioinformatics and molecular biology, where understanding the structure and function of proteins is critical. Classification, or supervised learning, is one of the major data mining processes. Pattern recognition involves assigning a label to a given input value. Protein classification is a problem of pattern recognition. Classification of protein sequences is an important tool for elucidating the structural and functional properties of newly discovered proteins. This protein classification is used in drug discovery, prediction of molecular functions, and medical diagnostics. Many techniques can be applied for classification tasks such as statistical techniques, decision trees, support vector machines and neural networks. Neural networks have been chosen as technical tools for protein sequence classification tasks because: features extracted from protein sequences are distributed in a high dimensional space and have complex features that can be satisfactorily determined using some parameterized approaches. And becomes difficult to model; And the rules produced by decision tree technology are complex and difficult to understand because properties are extracted from long character strings. In this paper, a comparative study of training feed forward neural networks is done using back propagation algorithm with three algorithms – back propagation algorithm, Levenberg Marquardt algorithm and genetic algorithm as optimizers. The efficiency of the three algorithms is measured in terms of convergence rate and performance accuracy.

***Keywords*:** Classification, Pattern, Input, Bioinformatics and Molecular Biology etc.

1.1 INTRODUCTION

Neural networks are simplified models of a biological neuron system; massively parallel distributed processing system composed of highly interconnected processing elements that have the capability to learn and thus acquire knowledge and make it available for use. , Various mechanisms exist to enable NNs to acquire knowledge [1]. In the training phase, neural networks extract the features of the input data. In the recognition phase, the network separates the patterns of the input data by features, and the recognition results are strongly influenced by the hidden layer [2]. Neural-network learning can be specified as a function approximation problem, where the goal is to learn an unknown function (or a good approximation of it) from a set of input-output pairs [3]. There is a need to develop an intelligent system to classify proteins that fall in a particular super family. Bioinformatics is the application of computer science and information technology to the fields of biology and medicine. Bioinformatics produces new knowledge as well as the computational tools to create such knowledge. The analysis and interpretation of biological sequence data is a fundamental task in bioinformatics. Classification and prediction techniques are one way to tackle such a task.
Data mining is a growing area of computer science. It is the process of finding new patterns in huge databases. Several algorithms have been proposed for the analysis of data. Algorithms analyze data and try to fit a model to the data. Sometimes KDD is another term used for data mining. KDD analyzes the data and extracts the necessary information and patterns from the analyzed data. Data mining uses algorithms to extract useful information and patterns in the data.

1.1.1 Neural networks have been chosen as an effective tool for protein sequence classification tasks because:
1) Features extracted from protein sequences are distributed in a high dimensional space with complex features that are difficult to model satisfactorily using some parameterized approaches.
2) The rules generated by the decision tree technique are complex and difficult to understand because attributes are extracted from long character strings [3].

1.2 BASIC CONCEPTS OF PROTEIN CLASSIFICATION IN ARTIFICIAL NEURAL NETWORKS

The study of protein classification in artificial neural networks is an exciting and specialized branch that is flourishing in bioinformatics and molecular biology. The main purpose of protein classification is to divide proteins into different classes, so that we can understand their structure, function, and various biological processes.

1.2.1 Here are some important concepts that are used in protein classification:
Structural Features: Various structural features are used to analyze the structure of proteins, such as the sequence of amino acids, amide bonds, and three-dimensional structure.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 01 January 2024

Functionality: Function of proteins is also an important classification parameter. This includes how the protein reacts, and what other molecules it may interact with.

Amino Acid Understanding (Amino Acid Composition): Studying the amino acid richness of proteins is also an important part. Some classification techniques take into account particular amino acid richness.
Machine Learning and Deep Learning: In modern neural networks, protein classification can also be done using machine learning and deep learning techniques. These techniques can learn from large and complex data sets and help classify new proteins.

Applications: The study of protein classification has applications in fields beyond children, medicine, biotechnology, and many more.

1.3 ARTIFICIAL NEURAL NETWORKS

Neural networks are simplified models of a biological neuron system, massively parallel distributed processing systems composed of highly interconnected processing elements that have the ability to learn and thereby acquire knowledge and make it available for use. Various mechanisms exist to enable NNs to acquire knowledge [1]. In the training phase, neural networks extract features from the input data. In the recognition stage, the network distinguishes the patterns of the input data by features, and the recognition result is greatly influenced by the hidden layer [2]. Neural-network learning can be specified as a function approximation problem where the goal is to learn an unknown function (or a good approximation of it) from a set of input-output pairs [3].

1.4 LITERATURE REVIEW

In [11], in contrast to this result, another study shows that the ADTree (alternative decision tree) algorithm provides the highest level of confidence to machine learning models. Additionally, a reliable colonic polyp detection system must ensure high sensitivity and specificity. Sensitivity measures the ratio between true positives (cases when a patient has a tumor and the system detects it, denoted by TP) and the total number of cancer patients. Specificity expresses the percentage of detecting true negatives (patients with non-cancer, denoted with TN). Thus, a specificity of 0.9 means that TN is correctly detected in 9 out of 10 cases. [14] Additionally, since professionals believe that sedentary lifestyle and Western diet are primarily responsible for the development of colorectal cancer, in this study, the environment in which the patients lived at the time of receiving the diagnosis. Yes, it was determined that this was taken into account. Of these, 12 qualitative data types are used: tumor status, T, N, M, Dukes classification, associated pathology, technical approach, complications, incidence, ultrasonography-dimensions as well as localization. There have been several studies on using artificial intelligence in computer-aided cancer detection software with the aim of reducing the risk of human error[10]. Decision tree (DT) is one of the most common traditional machine-learning algorithms used in medical applications for data analysis. Being one of the oldest and most excellent machine learning methods, decision trees have an architecture that is easy to understand and provides adequate results. Another machine learning technique that has recently gained popularity in cancer detection software is Support Vector Machines (SVMs). According to one study, SVM was used in detecting breast cancer (with an accuracy of 95%), multiple myeloma with an accuracy of 71%, and oral cancer with an accuracy of 75%.

1.5 OBJECTIVE

1. Using deep learning in cancer detection and finding out the most major differences between blood-work testing and genome-sequencing testing to detect leukemia
2. To find out the advantages and disadvantages of using different data modalities as input as well as the difference in accuracy of these two tests for results.

1.6 RESEARCH METHODOLOGY

Data mining is a growing field of computer science. It is the process of finding new patterns in huge databases. Many algorithms have been proposed to analyze the data. Algorithms analyze data and try to fit a model to the data. Knowledge search (KDD) is another term sometimes used for data mining in databases. KDD analyzes the data and extracts essential information and patterns from the analyzed data. Data mining uses algorithms to extract useful information and patterns in data.

1.7 EXPERIMENTAL RESULTS

The dataset was uploaded via a URL that contained the raw dataset in a text file. The filter() function made sure to remove all empty sequences and organized the data into a format that could be processed. The DNA sequence was transformed into a matrix. This was done with one hot encoding of Sklearn. LabelEncoder() turns the bases into an array of integers, and OneHotEncoder() turns the integer array into a matrix.
The dataset was split into training and testing with train_test_split() from the sklearn.model_selection function. The training set was further split into validation and training sets with validation_split = 0.25, which stores parts of the dataset to see whether dependent values are cancer or not predictive. The network architecture for this model was a 1D convolutional neural network. The model used the library Keras to easily build the network and implemented conv1d with filter=32 and kernel_size=12. 32

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 01 January 2024

filters were down-sampled in the pooling layer using MaxPooling1D. The matrix was generated from the individual pooling layer by replacing the columns in the next layer of the CNN. The Dense function had the activation function activation='relu' applied to the layer with 16 tensors, and the second activation function used was softmax. The dimensions of the images were reduced from 640x480 to 120x160 in order to train the model faster. The dataset was divided into training and test sets, and images for each type of WBC. Images were augmented to increase sample size and variation so that each training and test folder contained the same amount of images from different cell types. Now that the nucleotide has label-encoded values, this created a number sequence that can confuse the model. Each row corresponds to a nucleotide that has a predefined value that was written into the cells. Where A=Adenine , C=Cytosine, G=Guanine, T=Thymine

Table 1.1. DNA-sequence related labels and one-hot encodings

Nucleotide	One hot encoding			
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

This section presents classification reports on both the methods. The tables below display the forecast accuracy and total accuracy for each class. This is the overall performance of the whole method.
Table 1.2 shows the classification reports for the first method for genomic sequencing. The reported accuracy was similar to the accuracy in the confusion matrix plot. Class 0 is the positive cancer marker, and class 1 is for non-cancerous markers.

Table 1.2. Classification Report of Genomic Sequencing Method

Class	**Recall**	**Precision**	**F1-score**	**Total Accuracy**
0	0.95	0.97	0.97	0.97
1	0.97	0.95	0.97	

The lines begin to flatten around 0.97, which is around the same level where the model reached its highest accuracy - the result is compared to the confusion matrix True Positive, which was also 0.95, indicating that the plot is accurate. The interpretation of the values was that the model correctly labeled 95% of the cancer markers. The overall accuracy for the entire method is presented in the classification report and it had an accuracy rate of 97%. Despite the data being sampled from leukemia patients, the accuracy difference may be due to the 2000 rows of genome samples being sequenced and each row being treated as an input. The image sample size is 10000 images, which is five times the sample volume of the genome model taken. A larger dataset was needed for the test because reducing the number of images would not accurately represent the blood sample. Genomic methods showed that the two types of errors had lower values, but for the image processing confusion matrix, the values were different for each WBC type are neutrophils (0) made 422 false predictions, the highest among the four cell types. There are further questions about how well neutrophil images can be detected as a result of the high degree of false prediction of neutrophil levels.
The purpose of this method was to detect cancer markers on DNA sequences from cancer cells. In this test, a dataset containing 2000 lines of DNA sequences was used. Each row contains 50 nucleotides. The epochs to train the model were set to 50. The two figures below 10a–b show the performance of the model. Accuracy measures the prediction performance of the model, and model loss introduces the uncertainty of the model's prediction. The distance between the training and validation line is small in figure A and the accuracy plot. The training and validation lines start separating from each other at around 0.92 and stop at around 0.97. The confusion matrix showed that the model had a prediction score of 0.97 TP, meaning that it found the markers and correctly identified them at a rate of 97%. TN had a 99% rate of correctly predicting non-cancerous markers. The two error type classes had low percentages with 3% and 1%. The blood smear data samples were images of white blood cell subtypes that were part of the BCCD dataset. These samples are also in the BCCDs GitHub or Kaggle profiles. The data sample contains 10000 images in JPEG format which have been verified by experts. The WBC was colored to be more visible for the algorithm to recognize abnormal cells. It also included the cell-type labels in the CSV file, and in each folder, there were approximately 2500 augmented images of each cell-type.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 01 January 2024

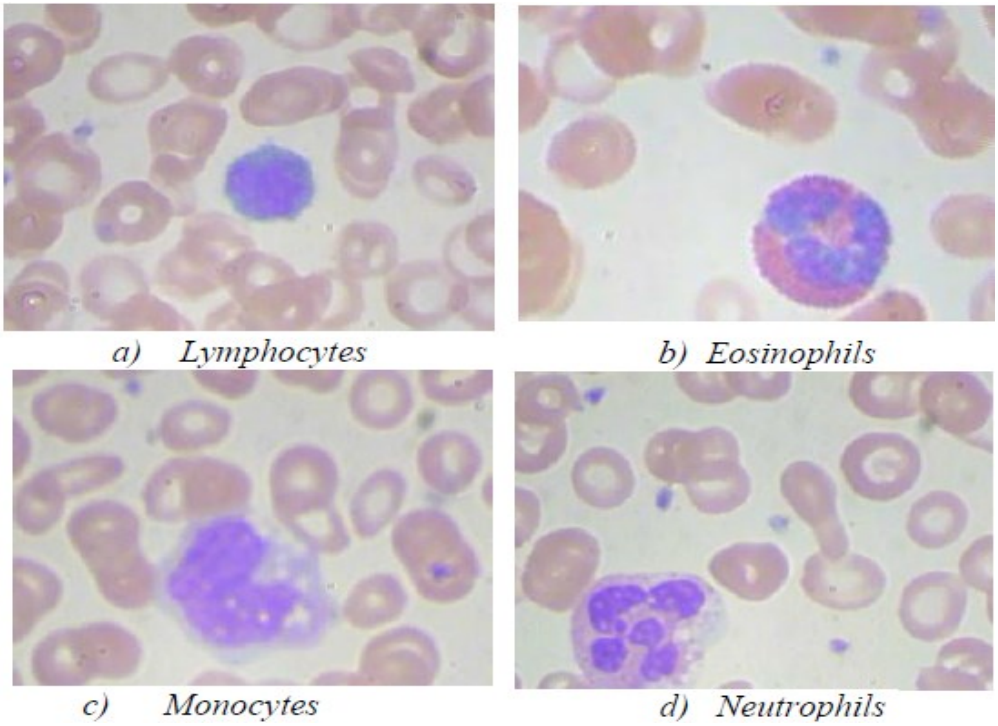

Figure 1.1. The picture shows microscopic images of different white blood cells.

1.9 CONCLUSION AND FUTURE WORK

In this paper, genomic sequencing and image processing methods can be applied to detect and predict leukemia in data samples. Further work in this area can be done using a single dataset using different neural network architectures. This is to see and compare which network algorithms can have better performance. A way to automate the pre-processing step could be devised for genomic sequences. This would contribute to the possibility of increasing the samples in the dataset and testing accuracy differences between methods. Various input samples have been interpreted and different data samples analyzed. The chapter discussed different effect sizes of data modalities to produce different results and resizing is not a good option for data samples. Furthermore, having limited access to data samples for blood smear images in comparison to DNA sequences limits the opportunity for the model to be tested multiple times with different inputs. These methods are automated versions of real-world techniques for detecting cancer. In this paper, genomic sequencing and image processing methods can be applied to detect and predict leukemia in data samples. Further work in this area can be done using a single dataset using different neural network architectures. This is to see and compare which network algorithms can have better performance. A way to automate the pre-processing step could be devised for genomic sequences. This would contribute to the possibility of increasing the samples in the dataset and testing accuracy differences between methods. Studying these concepts can help us develop new generation coding and classifying techniques for proteins, which can improve biologic experiments and lead to treatment with new and better drugs. The most important difference is that the genomic method is a binary classification. The image processing method is a multi-class classification, which is seen in the number of classes in the confusion matrix of the respective methods. Various input samples have been interpreted and different data samples analyzed. The chapter discussed different effect sizes of data modalities to produce different results and resizing is not a good option for data samples. Furthermore, having limited access to data samples for blood smear images in comparison to DNA sequences limits the opportunity for the model to be tested multiple times with different inputs. These methods are automated versions of real-world techniques for detecting cancer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 01 January 2024

REFERENCES

[1]. Cathy wu et. al., Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. Machine Learning Special issue on applications in molecular biology, 21(1-2):353–360, Nov.(1995).
[2]. Qicheng Ma and Jason T. L. Wang. Biological data mining using Bayesian neural networks: A case study. International Journal on Artificial Intelligence Tools, 8, 1993.
[3]. Dianhui Wang.G.: Protein sequence classification using extreme learning machine. In: IJCNN05, 3:1406–1411, 2005.
[4]. F.O. Karray and C. De Siva., Soft computing and Intelligent Systems Design, Theory, Tools and Applications. Pearson Education, 1st edition, 2009.
[5]. C. M. Bishop, Neural Networks for Pattern Recognition. Clarendon Press , Oxford, 1995
[6]. Satish Kumar. Neural Networks- A Classroom Approach. Tata McGraw-Hill
[7]. D. Wang and G. B. Huang, "Protein sequence classification using extreme learning machine," in Proceedings of International Joint Conference on Neural Networks(IJCNN,2005), Montreal, Canada, 2005.
[8]. Ali H., Sharif M., Yasmin M., Rehmani M.H., Riaz F. A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract. Artif. Intell. Rev. 2020;53:2635–2707. doi: 10.1007/s10462-019-09743-2.
[9]. Gheorghe G., Bungau S., Ilie M., Behl T., Vesa C.M., Brisc C., Bacalbasa N., Turi V., Costache R.S., Diaconu C.C. Early Diagnosis of Pancreatic Cancer: The Key for Survival. Diagnostics. 2020;10:869. doi: 10.3390/diagnostics10110869.
[10]. Nielsen M. Neural Networks and Deep Learning. [(accessed on 14 March 2021)]; Available online: http://neuralnetworksanddeeplearning.com.
[11]. Muhammad W., Hart G.R., Nartowt B., Farrell J.J., Johung K., Liang Y., Deng J. Pancreatic Cancer Prediction through an Artificial Neural Network. Front. Artif. Intell. 2019;5 doi: 10.3389/frai.2019.00002.
[12]. Chao W.-L., Manickavasagan H., Krishna S.G. Application of Artificial Intelligence in the Detection and Differentiation of Colon Polyps: A Technical Review for Physicians. Diagnostics. 2019;9:99. doi: 10.3390/diagnostics9030099.
[13]. Rwala P., Sunkara T., Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Prz. Gastroenterol. 2019;14:89–103. doi: 10.5114/pg.2018.81072.
[14]. Rampun A., Wang H., Scotney B., Morrow P., Zwiggelaar R. Classification of mammographic microcalcification clusters with machine learning confidence levels; Proceedings of the 14th International Workshop on Breast Imaging; Atlanta, GA, USA. 8–11 July 2018.
[15]. Goel N., Yadav A., Singh B.M. Medical image processing: A review; Proceedings of the IEEE Second International Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity (CIPECH); Ghaziabad, India. 18–19 November 2016.
[16]. Kourou K., Exarchos T.P., Exarchos K.P., Karamouzis M.V., Fotiadis D.I. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 2015;13:8–17. doi: 10.1016/j.csbj.2014.11.005.
[17]. C. M. Bishop, Neural Networks for Pattern Recognition. Clarendon Press , Oxford, 1995
[18]. Satish Kumar. Neural Networks- A Classroom Approach. Tata McGraw-Hill
[19]. D. Wang and G. B. Huang, "Protein sequence classification using extreme learning machine," in Proceedings of International Joint Conference on Neural Networks(IJCNN,2005), Montreal, Canada, 2005.
[20]. Cathy, michael berry, sailaja sivakumar etc..Neural networks for full time protein sequence classification: sequence encoding with singular value decomposition.Kluwer Academic publishers, 1995.
[21]. Edgardo A.Ferran, Pascual Ferrara etc...Protein classification using artificial neural networks.ISMB -93 proceedings, 1993.