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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

Classification of Protein Sequences for Cancer Diagnosis Using Artificial Neural Network

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

Abstract: **Classification or supervised learning, major data mining is taken as one of these processes. When we assign a label to the value of the given input in the recognition of a pattern. The problem of protein classification is the recognition of patterns. The classification of protein sequences is taken as an important tool to elucidate the structural and functional properties of newly discovered proteins. Protein classification is used for drug discovery and prediction of molecular functions and diagnostics in medicine for a variety of techniques on classification tasks. They are such statistical techniques, decision trees, support vector machines and neural networks and are thus taken as examples. The work is done using a feed forward neural network approach. Neural network as a technical tool is a technique chosen for protein sequence classification tasks, which can be distributed in higher dimensional space to facilitate the protein sequences. Using some parameterized approaches it is difficult to characterize and identify the pathway and model it. Rule created by decision tree technology is used to understand complex and difficult problem. In this paper, a comparative study has been done using three algorithms. Training Feed Forward Neural Network is done using Back Propagation Algorithm. Efficiency has been observed to converge convergence rate and display accuracy using back propagation algorithms, Levenberg Marquardt algorithms and genetic algorithms.**
Keywords: **Classification, Rule, Problem, Efficiency etc.**

INTRODUCTION

Computer science and information technology have an important place in the fields of bioinformatics, biology and medicine, which are applied in all fields. Bioinformatics is incomplete without the knowledge of computer science and information technology, new knowledge as well as computational tools are created for bioinformatics. Classification and prediction techniques are a way of dealing with such tasks in which the analysis and interpretation of biological sequence data is a fundamental task in bioinformatics. Along with the process of finding new patterns in huge databases, data mining is a very big field, we can say that in today's era, data mining computer science has made a lot of progress.

Protein secondary structure prediction (PSSP) is considered one of the major challenging tasks in bioinformatics, so several solutions have been proposed to address that problem through trying to obtain more accurate prediction results. The goal of this paper is to develop and implement an intelligent based system to predict the secondary structure of a protein from its primary amino acid sequence using five models of neural networks (NNs). These models are Feed Forward Neural Network (FNN), Learning Vector Quantization (LVQ), Probabilistic Neural Network (PNN), Convolutional Neural Network (CNN), and CNN Fine Tuning for PSSP. Two datasets have been used to evaluate our approach. The first contains 115 protein samples, and the second contains 1854 protein samples.

Almost all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and driven by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships among all proteins whose structures are known. As such, it provides a comprehensive survey of all known protein layers, detailed information about the close relatives of a particular protein, and a framework for future research and classification. Bioinformatics includes technology that uses computers to store, retrieve, manipulate and distribute information relating to biological macromolecules such as DNA, RNA, and proteins. The use of computers in genome mining is absolutely essential for information gathering and knowledge creation [1]. Protein structure prediction methods are classified under bioinformatics which is a broad field that combines many other fields and disciplines such as biology, biochemistry, information technology, statistics and mathematics [2].

LITERATURE REVIEW

[10] There have been several studies on the use of artificial intelligence in computer-assisted cancer detection software to reduce the risk of human error. Decision trees (DTs) are one of the most common machine learning methods used in medical data analysis systems. As one of the oldest and most advanced methods of machine learning, Decision Tree has a design that is easy to understand and provides great results. Another newly acquired machine learning technology in cancer detection software is Support Vector Machines (SVMs). According to one study, SVMs were used in the detection of breast cancer (95% accuracy, multiple myeloma with 71% accuracy, and oral cancer with 75% accuracy). In [11], in contrast to this effect, other studies show that the ADTree (another tree to decide) algorithm provides a higher level of confidence in machine learning models. Additionally, a reliable polyp polyp detection system should ensure high sensitivity and clarity. Sensitivity measures the ratio between actual discharge (cases in which a patient has a tumor and the system detects it, defined by TP) and the total number of cancer patients. The specification indicates the percentage of cancer screening (non-cancer patients, marked with TN). Therefore, the 0.9 specification means that TN

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

is correctly detected in 9 out of 10 cases. [14] In addition, as experts believe that sedentary lifestyles and Western diets are a major factor in the development of colorectal cancer, in this study, the area in which patients lived when they received the diagnosis was. Yes, it was decided and considered. Of these, 12 types of quality data are used: tumor status, T, N, M, Dukes category, associated pathology, methodology, complications, cases, events, ultrasonography size and local practice.

Research Questions
R1: Cancer detection using deep learning What are the opportunities and challenges as well as the major differences between blood-work testing and genome-sequencing testing for leukemia detection?
R2: What are the advantages and disadvantages of using different data methods such as input, and the difference in accuracy of these two outcome tests?

Research Method
Data mining is a growing area of computer science. It is a process of finding new patterns in large warehouses. Several algorithms have been suggested for data analysis. Algorithms analyze data and try to match the model to the data. Sometimes Knowledge Search in Database (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 from data.

Data mining techniques
Access to data mining in archives differs from conventional access in a number of ways:
Question: The question may not be stated correctly or may be well constructed. The data miner is probably not sure what he wants to see.
Data: The accessed data is usually a different version of the original operating system. The details will need to be refined and refined to better support the mining process.
Output: Output query output is probably not a subset of a website. Rather, it is the result of some analysis of content on a website.
Model: The purpose of the algorithm is to insert a model into the data.
Preferences: Some conditions should be used to fit one model over another.
Search: All algorithms require a specific data search method. Each created model can be predictable or descriptive in nature. The speculative model makes predictions about data values using known results from different data. Mining data model mining operations include segmentation, regression, time series analysis, and forecasting. Descriptive model identifies patterns or relationships in data. Compilation, summarizing, organizational rules and search sequences are often viewed as natural descriptions.

Feature Background Techniques
The principle of subjugation is that the mark of an object that is measured on the scale has a very similar value to objects in the same category, and is very different from objects in different categories. This leads to the idea of looking at certain irreversible features in irreversible input variables. After the features are released, some of the targets may or may not be relevant to the concept. Therefore in order to reduce complexity, it is necessary to remove all non-essential and non-essential features. The factors released in protein sequence are as follows: isoelectric point, molecular weight, atomic structure and amino acid length. These extracts act as an insertion in a neural network that is built to predict the large family intake protein that is part of it.

Learning system
Neural networks can be done in three ways. They are as follows:
• supervised learning
• supervised learning
• Reinforced Education

RESULT ANALYSIS

Blood smear pictures
The blood smear data samples were pictures of the type of white blood cells that were part of the BCCD database. These samples are in the BCCDs GitHub or Kaggle profile. The data sample contains 10000 images in certified professional JPEG format. The WBC is colored to make it more visible so that the algorithm can detect abnormal cells. It also added cell type labels to the CSV file, and in each folder, there were about 2500 advanced images for each cell type.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

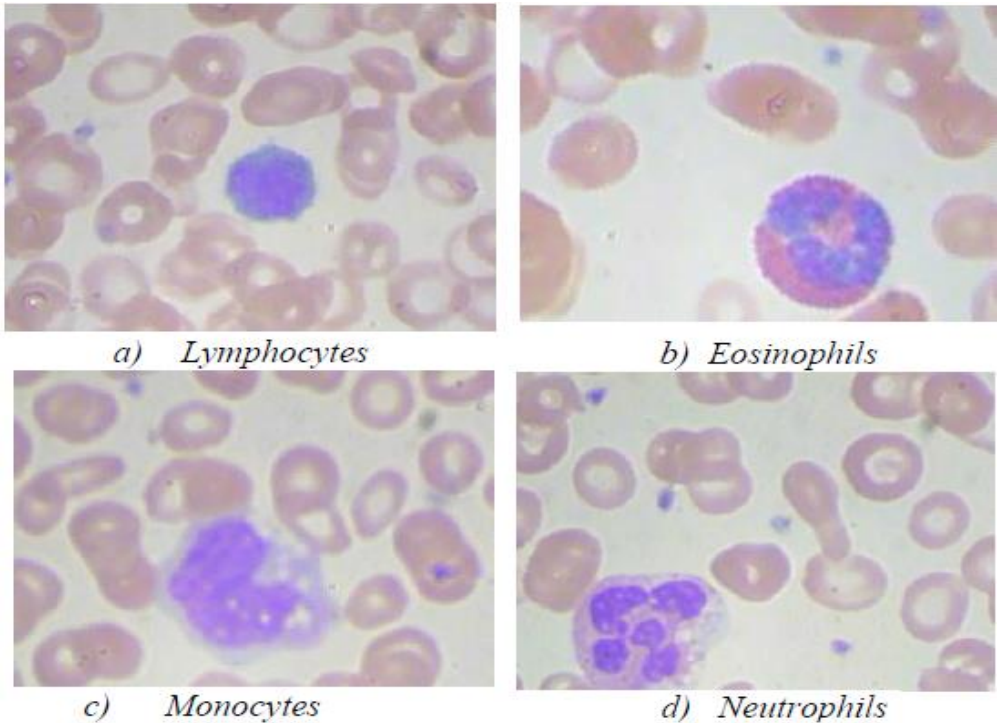

Figure 1.1. The image shows small images of different white blood cells.

Image size reduced from 640x480 to 120x160 for immediate model training. The database was divided into training sets and test sets, with images of each WBC type. Images are added to increase sample size and diversity so that each training and test folder has the same number of images from different cell types. Now that the nucleotide has labeled values, this creates a numerical sequence that can confuse the model. Each line corresponds to a nucleotide containing a predefined number that was recorded as a cell. When A = Adenine, C = Cytosine, G = Guanine, T = Thymine

Table 1.1. Labels related to DNA sequence and hot code

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

MODEL IMPLEMENTATION

The database is loaded with a URL containing the raw data in the text file. The filter function () ensures that you remove all empty spaces and organize the data into an unprocessed format. DNA sequences were transformed into a matrix. This is done with one hot Sklearn coding. LabelEncoder () turns the bases into whole numbers, and OneHotEncoder () turns a complete list into a matrix.

The database was split into training and testing with train_test_split () from the sklearn.model_selection function. The training set was further divided into validation sets and training sets with validation_split = 0.25, which retains parts of the database to determine whether dependent values are cancerous or unpredictable. The network structure of this model was the 1D convolutional neural network. The model used Keras library to easily build a network and used conv1d with filter = 32 and kernel_size = 12. 32 filters were sampled down to the composite layer using MaxPooling1D. The matrix is created from individual component layers by inserting columns in the next CNN layer. The dynamic function had the unlock function = 'relax' used in a 16-tensor layer, and the second activation function used was softmax.

After training the method, it uses the binary split function and introduces a learning curve to plan for network accuracy and loss. Prior to using the combination function, the model measured its losses using binary_crossentropy. The metaphor used for accuracy was the binary label _accuracy, which introduces the amount of speculation associated with dependent variations. For testing, Sklearn's model

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

EVALUATION

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%.

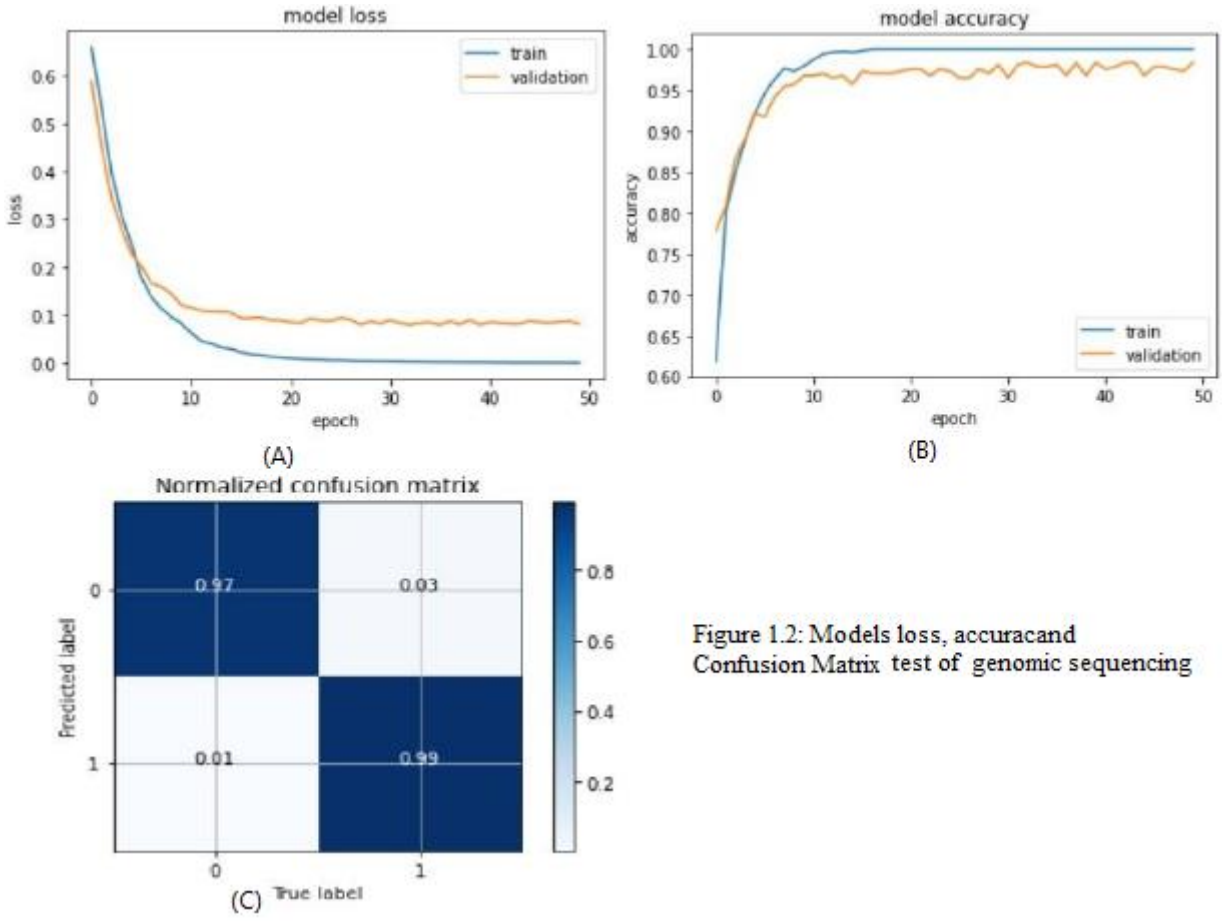

Figure 1.2: Models loss, accuracand Confusion Matrix test of genomic sequencing

REPORT OF CLASSIFICATION

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	
1	0.97	0.95	0.97	0.97

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

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 sampling the data 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 sample size of the image 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, with neutrophils (0) making 422 false predictions, the most among the four cell types was more. There are further questions about how well neutrophil images can be detected as a result of high levels of false prediction of neutrophil levels. According to the classification report, the percentage of recall is 61 and the accuracy is 88%. This result was mostly interpreted as the correct prediction, but with some results it was also observed that the lymphocyte classification report showed the opposite result with high recall but low accuracy, implying that the highest prediction method for class 1 Turned out to be wrong There can be a high percentage for both recall and precision. Difficulties arise in taking DNA samples, translating them into data and storing them in secure databases, which is a difficult task that requires resources and is expensive. If there are resources for DNA sequencing this would be the more optimal option.

CONCLUSION AND FUTURE WORK

Genomic method implementation for research questions 1 and 2 also provides a place to store DNA which is a potential substitute for DNA. WBC test is the best option which is very economical and we can control it manually. But the challenge before us is that it takes more time. The genomic method is known as a binary classification whose image processing method is a multi-class classification. has been analysed. It would be useful to discuss the different effect sizes of DNA as well as changing the size of the data samples to achieve different results. When access to data samples is limited in blood smear images compared to DNA sequences, the opportunity for the model to perform different tests using different inputs is limited. The automated version is used to detect cancer in the real world. In this paper, we used genomic sequencing and image processing methods to detect leukemia in data samples. We can do more work in the future with a single dataset using different neural network architectures. It can be compared to improve the performance of network algorithms. May contribute to amplifying samples and testing the difference in accuracy between their methods.

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
[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.