IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Cancer Care : Predicting the likelihood of cancer

[1]Mrs. Priya N, [2]Mr. Abhishek Esapnor, [3]Mr. Gagan Rao P, [4]Ms. Varshitha S

[1]Sr. Assistant Proffesor, [2]Student, [3]Student, [4]Student
[1]Department of ISE, New Horizon College of Engineering, Bangalore, India
[2]Department of ISE, New Horizon College of Engineering, Bangalore, India
[3]Department of ISE, New Horizon College of Engineering, Bangalore, India
[4]Department of ISE, New Horizon College of Engineering, Bangalore, India

Abstract : The goal of this research is to create a predictive model that analyzes a person's genetic makeup and forecasts their risk of acquiring cancer using deep learning algorithms. Through analysis of a person's genetic makeup, the discipline of genomics has demonstrated enormous potential in improving cancer detection and therapy. In order to reduce the amount of human intervention, our method compares several deep learning models that don't require feature engineering and gathers an original dataset. To make sure the gathered data is of high quality and appropriate for deep learning model training, pre-processing will be used. Using the pre-processed data, the predictive model will be trained and then assessed to determine how well it performs and how well it can generalize. The project's objectives are to increase patient outcomes, decrease the number of invasive screening techniques, and increase cancer detection rates. Treatment planning, tailored therapy, and cancer screening and diagnosis are some of the project's possible uses.

INTRODUCTION

One of the main causes of death worldwide is cancer, and the likelihood of a successful outcome can be greatly increased by early detection. Nevertheless, the accuracy of the current screening techniques is limited, and they can be costly and invasive. By analyzing a person's genetic makeup, the discipline of genomics has demonstrated enormous promise in terms of improving cancer detection and therapy. Our goal in this project is to create a predictive model that analyzes a person's genetic makeup and forecasts their risk of acquiring cancer using deep learning algorithms.

As In order to reduce the amount of human intervention, our method compares several deep learning models that don't require feature engineering and gathers an original dataset. We will gather clinical characteristics, genetic sequences, and other pertinent patient data, as well as genetic data pertaining to cancer. Pre-processing will be applied to the gathered data to guarantee its quality and suitability for deep learning model training. Using the preprocessed data, the predictive model will be trained and then tested to determine how well it performs and how well it can generalize. To understand the elements influencing the model's predictions, an analysis of the predictions made by the model will be conducted. The model can be used in the real world by being implemented in a production setting after it has been trained, assessed, validated, and interpreted.

The goal of this research is to create a predictive model that analyzes a person's genetic makeup and forecasts their risk of acquiring cancer using deep learning algorithms. Treatment planning, tailored therapy, and cancer screening and diagnosis are some of the project's possible uses.

NEED OF THE STUDY.

The goal of this project is to create a precise deep learning model that can use a person's DNA sequence to predict their risk of developing cancer. The algorithm will be trained on a dataset of DNA sequences and associated cancer outcomes in order to accomplish this. The objective is to produce a trustworthy prognostic tool that researchers and physicians may use to pinpoint people who are very susceptible to cancer. The incidence and death of cancer will eventually be decreased thanks to the use of this technology in individualised preventative and treatment plans..

RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study, Data and Sources of Data, study's variables and analytical framework. The details are as follows;

3.1 A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation

Literature survey is Several cancer forms, including breast, colon, head, kidney, lung, thyroid, and uterine cancer, are predicted using the degree of DNA methylation in areas designated as promoters and probes.To reduce high-dimensional and high-noisiness DNA methylation data, a hybridized strategy based on feature selection and extraction techniques is employed. The suggested hybridization approach's efficacy is assessed based on the precision of various classifiers, including SVM, Random Forest, and Naïve Base.

3.2 Genetic predisposition to prostate cancer: an update" by Holly Ni Raghallaigh and Rosalind Eeles

This paper's primary goal is to present a thorough summary of the genetic foundation of prostate cancer. GWAS has been used to find shared genetic variations linked to diseases and to search across a large number of people's whole genomes to find shared genetic variants. Using next-generation sequencing (NGS), researchers have been able to find rare and highly penetrant genetic variations linked to prostate cancer. Based on these genetic variants, Polygenic Risk Scores are then calculated, which quantify a person's genetic risk for a disease.

3.3 Predicting cancer tissue-of-origin by a machine learning method using DNA somatic mutation data

This work aims to give a thorough analysis of the use of a machine learning-based method to the prediction of cancer genesis tissue using somatic mutation data. Using somatic mutation data from 4,000 tumours covering 24 cancer types, scientists created a random forest classifier and found 220 useful somatic mutations to train it. They used independent test sets and cross-validation to assess performance. Their average tissue-of-origin prediction accuracy was 88%.

3.4 Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer

They created a deep learning-based technique to predict prostate cancer relapses. They made advantage of 400+ patients' multi-omics data (gene expression, DNA methylation, miRNA). They combined several omics to provide a thorough grasp of biological processes. They used survival analysis and cross-validation to assess the model's performance. Their relapse prediction accuracy was 76.6%. However, the DNN-based feature learning's interpretability still has to be improved.

3.5 Pathway Analysis of Marker Genes for Leukemia Cancer using Enhanced Genetic Algorithm-Neural Network (enGANN)

Finding physiologically relevant and computationally efficient analytic models to examine the gene-to-gene interactions underlying a given disease's pathophysiology is the primary goal. First, genetic interactions between genes were modeled using the enhanced genetic algorithm-neural network (enGANN), a suggested inference technique, to evaluate the interactions between genes and their related biological functions. After that, the modelled interaction network was shown using Cytoscape, and the STRING online protein interaction database was used to confirm any potential marker routes. This methodology captures the differences between cancer and normal samples through Network Analysis of gene-gene interactions. They did not, however, look at how having different cancer subtypes would affect the inference performance.

3.6 A Multi-Learning Training Approach for distinguishing low and high-risk cancer patients.

This research paper's primary goal is to develop a Multi Learning Training algorithm that uses supervised, unsupervised, and self-supervised learning techniques to identify low- and high-risk cancer patients by utilizing the interaction of clinical and molecular features. The model applies feature selection to clinical, molecular, and therapeutic data as input. In order to produce gene expressed-based signatures, the Feature Extraction module is designed to perform Patient Clustering, Gene Clustering, and Gene Denoising.Finally, a supervised ensemble learning technique is employed by the Ensemble Predictor component to categorize cancer patients into risk groups.

3.7 A convolutional neural network-based ensemble method for cancer prediction using DNA methylation data.

One of the main goals of this study is to offer an ensemble approach based on convolutional neural networks to investigate the internal links between DNA methylation and cancer. A t-test is used in feature selection to identify methylation locations that are significantly different between normal and malignant samples. In the initial phases, cancer samples are separated from normal samples using classification techniques such as the Naive Bayesian Classifier, k-Nearest Neighbour, Decision Tree, Random Forest, and Gradient Boosting Decision Tree. Because of its capacity to process data from biological visual systems, an ensemble model based on convolutional neural networks is developed in later stages. The primary benefit is that the multi-model ensemble method based on convolutional neural networks can automatically learn the complex relationships between the classifiers and produce improved prediction results. Since supervised learning models have been utilized, careful feature selection is necessary prior to classification.

3.8 Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method

Numerous models, including the Generalized Linear Model (GLM), K-Nearest Neighbor (KNN) Classifier, Support Vector Machine (SVM) Classifier, Naïve Bayes Classifier, and C5.0 Decision Tree Classifier, comprised the ensemble structure. Rather than utilizing an ensemble directly, the research examines the performance of several models on an individual basis. It investigates the Hepatitis B virus's DNA sequence. Although there is a substantial association between Hepatitis B and Liver Cancer, this method is not accurate in predicting other cancers that may not have a comparable condition, nor in people who do not have Hepatitis B.

3.9 Analysis of Prostate Cancer DNA Sequences Using Bi-direction Long Short Term Memory Model.

This work attempts to predict cancer using a hybridized method based on feature extraction and selection procedures by utilizing the degree of DNA methylation in promoter and probe regions. The recommended method overcomes the high-dimensionality issue with the DNA methylation data by utilizing a filter feature selection technique known as (F-score). The study contrasts several models, such as support vector machines, Random Forest, and Naive Bayes. It contrasts the accuracy and Fmeasure of each of these models. The primary benefit is that there is no feature engineering or feature selection done. The neural network is given the raw data and the learning and predicting is completely left up to the neural network. However, the paper completely ignores the intron regions of the DNA. Although introns do not directly encode proteins, they play important roles in gene regulation and can influence the expression of the gene. Hence introns can affect protein synthesis and do play a role in DNA damage. The paper extracts just the exon regions of the DNA for training and predicting.

IV. RESULTS AND DISCUSSION

4.1 Model Training

After Data Pre-Processing, Deep learning algorithms are utilized to train the predictive models. These models are trained with the training set through several epochs, where the optimization algorithm and loss function are utilized to adjust its parameters.

4.1.1 Hybrid CNN RNN Model

This model is made up of two prevalent deep learning models. The first layer is the Input Layer which accepts the gene sequences.Predicting likelihood of Cancer, based on gene markers 29.

- The CNN layer is made up of 2 parts, a Convolution layer and a Maxpooling layer. The second part is the RNN layer which utilizes LSTM cells(Long Short-Term Memory) for sequential pattern recognition.
- Next, we combined both the CNN and RNN components using concatenate function provided by Tensorflow.
- The output layer is made up of just one neuron that uses sigmoid activation function. The model is compiled with Adam optimizer.

4.1.2 LSTM Model

This model consists of 4 layers, input layer, output layer, and two hidden layers in between. Like the previous model, the input layer accepts gene sequences as input and forwards it to the hidden layer. The two hidden layers are made up of 64 LSTM cells each. The output layer is made up of a single dense neuron that uses sigmoid activation function. Adam optimizer was used to compile the model.

4.1.3 Bi- LSTM Model

The Bi-LSTM model is compiled the same way as LSTM model, with the exception being, the LSTM cells used in the two hidden layers are made bi-directional. This change, however, did not result in a huge impact.

4.1.4 k-NN (without PCA)

The k-NN (k-Nearest Neighbors) model employed in this context was characterized by its simplicity, with a parameter setting of k, representing the number of nearest neighbors, set to 5. This model was directly fitted with the preprocessed data, emphasizing a straightforward approach to pattern recognition and classification. The essence of the k-NN algorithm lies in its reliance on the proximity of data points in feature space, where predictions are determined by the consensus of the k closest neighbors.

4.1.5 k-NN(with PCA)

In enhancing the k-NN model, Principal Component Analysis (PCA) was incorporated to reduce the dimensionality of the feature space, mitigating the curse of dimensionality and potentially improving model performance. By capturing the most significant variations in the data through orthogonal transformations, PCA allowed for a more efficient representation of input features. The amalgamation of k-NN with PCA aimed to strike a balance between computational efficiency and preserving essential information, contributing to a more streamlined and effective.

4.2 Performance Evaluation:

Statistical Overview and table of performance of the trained model are as shown :

Cancer Types	Model Used				
	Hybrid CNN RNN	Bi- LSTM	LSTM	KNN(with PCA)	KNN(without PCA)
Colon	70.79	70.79	69.32	96.02	99.5
Lymphoma	71.20	71.20	70.00	99.20	100
Lung	72.19	72.19	70.89	87.00	96.79
Uterus	85.19	89.50	89.50	100	100

Blood	78.92	64.64	66.14	95.71	99.64
Pancreas	88.54	86.63	76.63	97.45	100
Brain	67.48	67.475	68.89	98.54	99.5
Stomach	86.25	83.56	80.36	96.25	96.87
Bladder	86.25	82.75	82.79	100	100
Misc	83.22	80.22	77.00	98.89	100
Breast	64.74	61.73	62.55	93.91	97.39
Prostate	85.54	85.54	68.33	99.39	100
Kidney	88.05	88.05	85.00	90.56	96.85
Liver	89.24	89.24	89.24	98.73	98.73
Throat	87.73	87.73	85.33	94.47	98.15

Table 4.1 Performance Analysis of Different Models on Each Cancer type

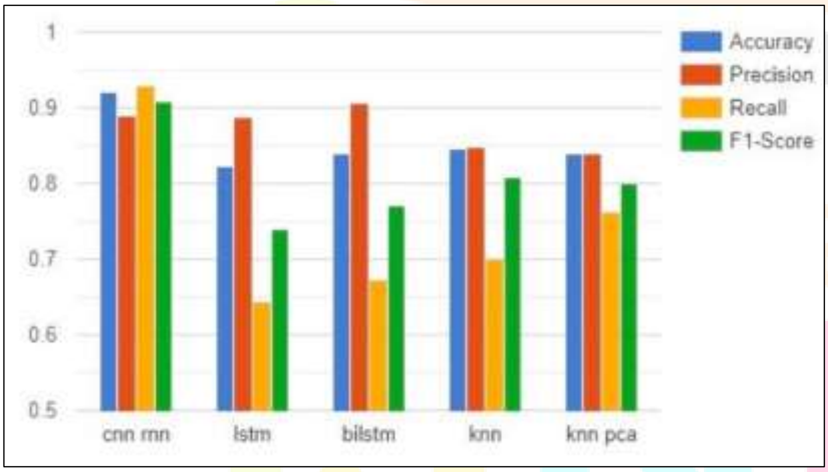

Fig 4.1 Overall Analysis of the models

References

[1] Raweh, A. A., Nassef, M., & Badr, A. (2018). A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation. IEEE Access, 6, 15212-15223.
[2] Ni Raghallaigh, H., & Eeles, R. (2022). Genetic predisposition to prostate cancer: an update. Familial Cancer, 21(1), 101-114
[3] Liu, X., Li, L., Peng, L., Wang, B., Lang, J., Lu, Q., ... & Zhou, L. (2020). Predicting cancer tissue-of-origin by a machine learning method using DNA somatic mutation data. Frontiers in genetics, 11, 674.
[4] Wei, Z., Han, D., Zhang, C., Wang, S., Liu, J., Chao, F., ... & Chen, G. (2022). Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer. Frontiers in Oncology, 12.
[5] Wong, H. C., Lee, C. S. K., & Tong, D. L. (2018, October). Pathway Analysis of Marker Genes for Leukemia Cancer using Enhanced Genetic Algorithm-Neural Network (enGANN). In 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 118-121). IEEE.
[6] Povoa, L. V., Calvi, U. C. B., Lorena, A. C., Ribeiro, C. H. C., & Da Silva, I. T. (2021). A Multi- Learning Training Approach for distinguishing low and high risk cancer patients. IEEE Access, 9,
115453-115465.Dept. of CSE June – November, 2023 43

[7] Xia, C., Xiao, Y., Wu, J., Zhao, X., & Li, H. (2019, February). A convolutional neural network based ensemble method for cancer prediction using DNA methylation data. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (pp. 191-196).

[8] Muflikhah, L., Widodo, N., & Mahmudy, W. F. (2020, December). Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method. In 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 37-41). IEEE.

[9] Abass, Y. A., Adeshina, S. A., Agwu, N. N., & Boukar, M. M. (2021, November). Analysis of Prostate Cancer DNA Sequences Using Bi-direction Long Short Term Memory Model. In 2021 16th International Conference on Electronics Computer and Computation (ICECCO) (pp. 1-6). IEEE.