UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume: 6 | Year: 2024 | ISBN No. : 2976-2278

Hepatocellular Carcinoma Prediction Using Deep Learning Model Based On DNA Methylation

Aminul Hakim MD Rafie, Weng Howe Chan
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia
aminulhakim.mr@graduate.utm.my, cwenghowe@utm.my

Abstract— **The study regarding the relationship between Hepatocellular Carcinoma (HCC) prediction and DNA methylation data is still new currently. Besides that, the large scale of genomic dataset has become a challenge to current machine learning model prediction in term of time complexity and accuracy. The deep learning model is suggested to be an alternative to the machine learning method to deal with the high-dimensional DNA methylation dataset. Thus, the purpose of the research is to predict HCC liver cancer disease by implementing a deep learning model prediction based on DNA methylation biomarkers. From that, three research objectives were identified in this research. Firstly, previous machine learning and deep learning model implementation in HCC liver cancer disease prediction will be studied. Second, a predictive model for HCC liver cancer disease will be developed using deep learning architecture. Lastly, the performance of the proposed model was measured and evaluated based on its accuracy, sensitivity, and specificity. In this research, a Deep Neural Network (DNN) model will be employed consisting of multiple hidden layers to learn the pattern of HCC and normal samples from the DNA methylation data input. DNA methylation microarray dataset, GSE113017 will be taken from Gene Expression Omnibus (GEO) database. The dataset comprises 60 files of HCC and normal tissue samples. The result of this study showed the DNN model achieved an accuracy of 95%, sensitivity of 90%, and specificity of 100%. In summary, the DNN model outperformed the other machine learning models in predicting HCC disease based on DNA methylation dataset.**

Keywords – Hepatocellular Carcinoma, DNA methylation, machine learning, deep learning, model prediction

I. INTRODUCTION

Hepatocellular carcinoma (HCC) is the most prevalent type of liver cancer that is considered a high mortality rate cancer. HCC is the fifth global major cancer and the third leading cause of cancer death [1]. The prediction of HCC survival is substandard overall as the survival rate in a 5-year relative is only 18.4% [2]. Previously, many researchers have propounded machine learning and deep learning predictive models to predict HCC disease in their studies. However, the model is prone to overfitting with the large-scale dataset [3]. On the other hand, methyl-based biomarkers are discovered to have potential to be useful information for the HCC disease diagnosis process [4]. The study of the relationship between HCC disease and DNA methylation is in the commencement stage which includes only a few researches currently. Meanwhile, the research on HCC disease prediction via deep learning on DNA methylation has not been found [5].

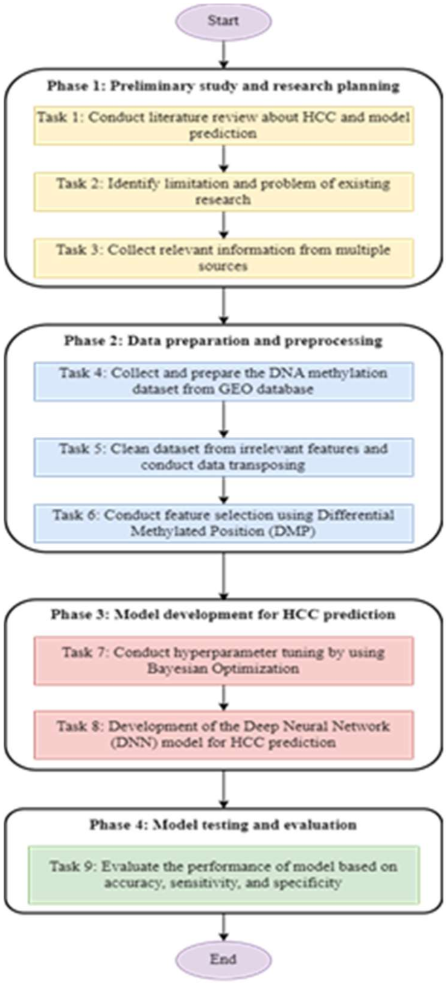
The high dimensionality of genomic data has become a challenge to machine learning approaches since it leads to computational inefficiency and overfitting [6]. Therefore, the alternative of the machine learning method with high-efficiency output needs to be identified in predicting HCC disease using DNA methylation data. This study constructs an aim to implement a deep learning model with optimized parameters in predicting HCC liver cancer disease using DNA methylation biomarkers. The objectives of the research are: (1) To study the previous machine learning and deep learning model implementation in HCC liver cancer disease prediction. (2) To develop a predictive model on HCC liver cancer disease using Deep Neural Network (DNN) with Bayesian Optimization. (3) To measure and evaluate the performance of the proposed deep learning model for HCC prediction based on the result of accuracy, sensitivity, and specificity. In short, this research will investigate the potential of deep learning methods as an alternative to machine learning in improving prediction performance on a large-scale dataset.

II. LITERATURE REVIEW

Primary liver cancer and secondary or metastatic liver cancer are the two categories of liver cancer. The most common primary liver cancer type is HCC which starts from the main type of liver cell named hepatocyte. DNA methylation is defined as a methyl group adjunction into the DNA strand and usually attaches at the fifth carbon atom of the cytosine ring. Many HCC predictive models have been developed by using machine learning architecture in previous research. For example, there is a study about the relationship between the overall survival of HCC patients and DNA methylation using SVM [3]. The survivability prediction is based on three categories; high risk, intermediate-risk, and low risk.

Next, research about prediction of the HCC disease based on 450K DNA methylation data using ANN [5]. Differentially expressed genes are applied to attain an improvement of model performance and indirect high-dimensional data processing. However, the model only comprises one hidden layer and is not suitable for the more complex computation. Research on HCC risk development prediction among the HCV with advanced fibrosis patients using decision trees has also been done [7]. However, the DT algorithm needs a boosting before it can perform well.

III METHODOLOGY


Start
Phase 1: Preliminary study and research planning
Task 1: Conduct literature review about HCC and model prediction
Task 2: Identify limitation and problem of existing research
Task 3: Collect relevant information from multiple sources
Phase 2: Data preparation and preprocessing
Task 4: Collect and prepare the DNA methylation dataset from GEO database
Task 5: Clean dataset from irrelevant features and conduct data transposing
Task 6: Conduct feature selection using Differential Methylated Position (DMP)
Phase 3: Model development for HCC prediction
Task 7: Conduct hyperparameter tuning by using Bayesian Optimization
Task 8: Development of the Deep Neural Network (DNN) model for HCC prediction
Phase 4: Model testing and evaluation
Task 9: Evaluate the performance of model based on accuracy, sensitivity, and specificity
End

Research Framework

A. Data Preparation

The raw dataset of HCC DNA methylation microarray data is retrieved from the GEO database. The GEO accession number of the data is GSE113017. The data file that is downloaded in the .txt format contains 30 pairs of HCC and adjacent liver controls that have been methylated and unmethylated. From the txt file, the dataset is exported into a csv file to ease the data preprocessing task in the next step. There are 485577 rows and 181 columns in this raw DNA methylation data

B. Data Cleaning

The unmethylated signal matrix, methylated signal matrix, and detection p-value are three kinds of data that are provided for each HCC and normal sample. The methylated signal matrix data of the CpG sites bring a significant value to the model prediction. Therefore, the unmethylated signal matrix and detection p-value columns for each sample are removed to reduce the number of columns to 61. The methylated value is significant in methylation differential analysis in improving further statistical validity [8].

C. Data Transposing

By transposing the cleaned dataset, the position of rows and columns dataset has been twisted to make CpG sites the new columns. It means the current rows of the dataset are the HCC and normal samples with their methylated signal scores in each CpG site. Lastly, a column of the class label is appended to the table.

D. Differentially Methylated Position (DMP)

The dimension of the dataset is scaled-down since this high features data will cause overfitting and high gradients variance when being fitted to the model prediction directly [8]. However, common feature selection techniques including the principal component analysis (PCA), Lasso, or Relief-F are not appropriate for DNA methylation microarray data as the correlation between features dataset are not appraised correctly. Therefore, the biological meaning in the data will be affected when the features are shrunk by those techniques.

The DMP is used in this research as it links the CpG sites to the genes that are associated with cell attachment and vital to organ development [9]. In this research, the CpG sites or dataset features are filtered using the Methylize package based on the DMPs with FDR scores below 0.01. It aggregates with the Methylprep and Methylcheck python packages to calculate the DMP of the CpG sites using their methylated signal value. The DMPs are calculated via a regression method between methylation signal value and phenotype data of the sample dataset. The logistic regression is implemented since the sample contains only two options of phenotype data (HCC or normal). As a result, there 41923 CpG sites remain in the dataset. Lastly, each sample of the dataset is labeled either as HCC or normal accordingly before it can be fitted as input data for model prediction.

E. Deep Neural Network Model Development

The DNN model consists of three connected input, hidden, and output layers. Each layer will be equipped with a specific parameter, activation function, and type of layer. The optimal DNN model parameter value is determined via Bayesian hyperparameter tuning process. Bayesian optimization reveals the optimal black-box function in the parameter in the model [8]. The parameter value in the model will be determined based on the probabilistic result of the algorithm. Four parameters are tuned including number of hidden layers that comprise of dense and dropout layer (1-5), the number of layer nodes (16-128), the rate of dropout (0.0-0.5), and the amount of learning rate (0.0001-0.1). As a result, the best value for each hyperparameter is 3 hidden layers, 91 nodes, 0.0367 of dropout rate, and 0.0032 of learning rate.

Rectified Linear Unit (ReLu) activation function is used in the input and hidden layer of the model prediction. The ReLu function ensures the layer returns the positive value from the input received so that model will be trained easier and promise a better result. However, the output layer applies a sigmoid function as it guarantees the production of the prediction model is always between 0 (normal) and 1 (HCC). The model compilation is the final stage to develop the model before training begins. In this research, the model compilation used the binary cross entropy loss function as it examines the variance between the predicted probability distribution and the actual binary class [10]. Next, the Adagrad optimizer with a 0.0244 learning rate is used in this model since it provides high processing speed with the best prediction accuracy [11].

The DNN model prediction is trained and evaluated iteratively via 5-fold cross-validation. This study used 32 batch sizes of samples to fit in one epoch. Besides that, the model training used 50 epochs to learn the dataset. For each iteration, the performance of the model is evaluated based on the accuracy, sensitivity, and specificity metrics. Next, the mean score of each performance metric is calculated to determine the overall performance of the model. The calculation formula of each performance metric is as follow:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \tag{1}$$

$$\text{Sensitivity} = TP / (TP+FN) \tag{2}$$

$$\text{Specificity} = TN / (TN+FP) \tag{3}$$

IV. RESULT AND DISCUSSION

By using the developed DNN model, the model performance is evaluated based on three performance metrics; accuracy, specificity, and sensitivity. Besides that, a comparative analysis based on the obtained results is conducted between the DNN model and other machine learning techniques.

A. Confusion Matrix

In this study, the value of TP represents the number of patients that are identified correctly to have HCC and need treatment. On the other hand, the TN value shows the total number of patients who are healthy and certainly do not require any treatment. The FP or type I error represents the number of wrong-predicted HCC patients that are actually healthy from the disease. Meanwhile, FN, the type II error, portrays the number of patients that are suffering from HCC to be in the incorrect prediction as a healthy person.

Based on Table I, 30 samples of healthy patients are correctly identified as the normal class sample. It shows the model successfully predicted all the actual healthy patients as the normal class. Besides that, 9 out of 10 HCC patients are diagnosed accurately as the HCC class by the model prediction. Moreover, the confusion matrix reported there is no type I error from the model prediction. It means the model performed very well in evaluating normal or healthy patients. However, there are 3 patients who are actually experiencing the HCC disease and were predicted to be healthy patients.

CONFUSION MATRIX

	Predicted HCC	**Predicted Normal**
Actual HCC	*27*	*3*
Actual Normal	0	30

B.Accuracy, Sensitivity, and Specificity

Table II shows both ANN models in Zhang's work and the proposed method have almost comparable performance overall. Zhang's ANN model prediction scored a sensitivity of 96.8%, better than the DNN model with 93.3%. However, the developed model of this study possessed a better specificity score (100%) than Zhang's ANN model (93.3%). In terms of accuracy, both models have equivalent scores with only 0.1% differences.

COMPARISON BETWEEN MODELS

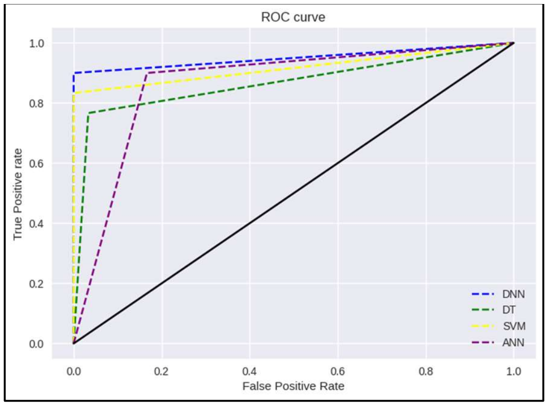
Works	**Methods**	**Accuracy**	**Sensitivity**	**Specitifity**
Zhang's work	ANN	95.1	96.8	93.0
This work	DNN (Proposed method)	95.0	90.0	100.0
	DT	86.7	76.7	96.7
	SVM	91.7	83.3	100.0
	ANN	86.7	90.0	83.3

After 5-fold cross-validation, the DNN prediction model performance comprehensively surpassed the other methods covered in this study. The DNN model scored a top performance in accuracy metrics with 95.0%, followed by the SVM (91.7%), DT (86.7%), and ANN (86.7%) models. Meanwhile, in sensitivity, the DNN shared a similarly high score with ANN since both had a sensitivity of 90%. The lowest sensitivity score among the models was the DT model since it obtained a sensitivity of 76.7%. SVM was in the middle rank, achieving an 83.3% score. However, the specificity of the SVM model showed a satisfactory score in contrast to its sensitivity score. Both DNN and SVM models were splendid in predicting the negative instances (Normal class) as they acquired a 100% of specificity score. Besides that, the DT model presents a better specificity score than ANN. The DT model scored a specificity of 96.7%; on the contrary, the ANN attained 83.3% only.

C. AUC-ROC

Based on the ROC generated, the Area Under the Curve (AUC) was calculated to measure the ability of each model to distinguish label class. Table III showed the DNN model had a better measure of separability since its AUC was the closest to

1. Besides that, the curve of the DNN model's ROC was the closest to the top-left corner of the ROC graph, thus demonstrating it was the better model in performance and accuracy.



AUC-ROC curve between models

COMPARISON BETWEEN MODELS

Model	AUC
DNN (Proposed method)	0.95
DT	0.8666666666666667
SVM	0.9166666666666667
ANN	0.8666666666666666

D. Discussion

The GSE113017 dataset from the GEO database was a DNA methylation microarray data that consisted of 30 pairs of HCC and adjacent liver control samples. The dataset was in balanced class distribution since it had a complement amount between positive instances and negative instances. The advantage of using the balanced dataset was the model could avoid bias and overfitting during the training. It was because the skewed class distribution of an imbalanced dataset would affect the prediction process and lead to poor performance.

On the other hand, an enormous amount of features in the DNA methylation data increased the dimensionality of the dataset. A high-dimension dataset tended to escalate the model complexity. Besides that, the noise data presented in the dataset can impact the model during training. As a consequence, the model consumed a long time during the training and produced low-accuracy results. The DMP was a feature selection approach that was specifically used to screen out the low-correlated CpG sites from the DNA methylation dataset. By implementing the DMP, the model could be trained with the most significant features. Hence, it reduced the possibility of overfitting and improved the performance of the model prediction.

Hyperparameters tuning was a fundamental aspect when developing prediction models. Good hyperparameters could maintain the behaviors of the training model and deliver a remarkable impact on the model's performance. Based on the Bayesian optimization, the developed DNN model acquired six hidden layers to learn the DNA methylation data pattern. 3 dense layers and 3 dropout layers are the optimal numbers of layers to ensure the model performs. The optimal number of hidden layers would minimize the time complexity while improving the performance accuracy (Uzair & Jamil, 2020). Besides that, the low dropout rate was to improve the model on regularization performance without slowing its convergence rate. Overall, the optimization of the model hyperparameter was parallel to the sample size and complexity of the model.

Comprehensively, the developed DNN model exhibited an outstanding performance in HCC disease prediction based on DNA methylation data. A high accuracy score illustrated the good capability of the prediction model to differentiate between HCC and healthy patients from their DNA methylation data. Besides that, the specificity of 100% indicated the prediction model learned the Normal class samples very well.

From the conducted analysis, the DNN model scored better performance than other methods based on the accuracy, specificity, and sensitivity metrics. Two prediction models, DNN and SVM, correctly predicted the whole negative instances in the dataset. Well-processed data during cleaning and feature selection might be one of the factors that influenced the high score of the models in specificity. However, in this study, the sensitivity metric was more significant than the specificity. The sensitivity performance of each model was evaluated to determine the best model prediction for HCC. A model with better sensitivity can detect potential HCC patients earlier and more accurately. The neural network models, DNN and ANN, scored the highest in sensitivity during the model evaluation.

V. CONCLUSION

This thesis concludes the outcomes from the study of HCC prediction based on DNA methylation using a deep learning model. Moreover, the objectives and scopes constructed in this study were successfully fulfilled. Besides that, research constraints and suggestions were discussed to provide assistance for improvement in the future.

A. Achievements

The study discovered the utilization of DNA methylation data in predicting HCC disease based on machine learning methods still in the infancy stage. Apart from that, the competency of the deep learning model in handling large and complex data from the previous works showed its potential to better learn the pattern of DNA methylation microarray data in predicting HCC disease. In short, the first objective of this study was achieved during the preliminary study stage.

Next, the second objective was completed during the process to develop the DNN prediction model with the optimization of the Bayesian algorithm. A deep understanding of the DNN model was achieved to design a pragmatic model architecture for disease prediction. Before that, the DMP method was used to filter the low correlated features from the dataset. The filtered features were used as the input during the training. Four hyperparameters, such as the number of neurons, hidden layers, dropout rate, and learning rate, were successfully tuned based on the developed Bayesian optimization method. After that, the study designed and constructed the DNN model with the value of the optimized hyperparameter. As a result, the developed model managed to obtain the prediction probability of the HCC disease using the DNA methylation dataset.

The developed DNN model was tested and evaluated based on three performance metrics to achieve the third objective. The model succeeded during the evaluation by securing excellent results in terms of accuracy, sensitivity, and specificity. In addition, the developed model outperformed the other machine-learning models in this study. Based on the performance evaluation, the developed DNN model was capable of predicting HCC disease patients based on the DNA methylation microarray dataset very well.

B. Future Works

It was suggested to conduct the study on the larger DNA methylation dataset since it helped the DNN model to learn the data point pattern better. With the large dataset, the DNN model could understand more about the variety of HCC patterns in the dataset. On the other hand, another advanced deep learning model such as the Multilayer Perceptrons, RNN, and many more also could be applied to evaluate the performance of different deep learning models in predicting HCC based on DNA methylation data.

ACKNOWLEDGMENT

The Universiti Teknologi Malaysia (UTM) supported this work through UTM Encouragement Research Grant with Cost Centre No. Q.J130000.3828.42J11.

REFERENCES

[1] El-Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology [Internet]. 2007 Jun 1;132(7):2557–76. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17570226

[2] Cicalese L. Hepatocellular Carcinoma (HCC): Practice Essentials, Anatomy, Pathophysiology. eMedicine [Internet]. 2022 Dec 22; Available from: https://emedicine.medscape.com/article/197319-overview#a2

[3] Dong RG, Yang X, Zhang XY, Gao P, Ke AW, Sun H, et al. Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. Cellular and Molecular Medicine. 2019 Feb 19;23(5):3369–74.

[4] Cheng J, Wei D, Ji Y, Chen L, Yang L, Li G, et al. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Medicine. 2018 May 30;10(1).

[5] Zhang Y, Ren Fangtao, Liu X, Zhang F. Prediction of Hepatocellular Carcinoma Diseases Based on Methylation Data and Screening of Hub Genes. CSAI '21: Proceedings of the 2021 5th International Conference on Computer Science and Artificial Intelligence. 2021 Dec 4

[6] Wu Q, Boueiz A, Bozkurt A, Masoomi A, Wang A, DeMeo DL, et al. Deep Learning Methods for Predicting Disease Status Using Genomic Data. Journal of biometrics & biostatistics [Internet]. 2018 [cited 2023 Jun 30];9(5):417. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6530791/#:~:text=Deep%20learning%20has%20shown%20breakthrough

[7] Hashem S, ElHefnawi M, Habashy S, El-Adawy M, Esmat G, Elakel W, et al. Machine Learning Prediction Models for Diagnosing Hepatocellular Carcinoma with HCV-related Chronic Liver Disease. Computer Methods and Programs in Biomedicine. 2020 Nov;196:105551

[8] Park C, Ha J, Park S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. Expert Systems with Applications. 2020 Feb;140:112873

[9] Spindola L, Santoro ML, Pedro Mario Pan, Vanessa Kiyomi Ota, Xavier G, Carvalho C, et al. Detecting multiple differentially methylated CpG sites and regions related to dimensional psychopathology in youths. 2019 Oct 21;11(1)

[10] Saxena S. Binary Cross Entropy/Log Loss for Binary Classification [Internet]. Analytics Vidhya. 2021 [cited 2023 Jun 30]. Available from: https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification

[11] Halgamuge MN, Daminda E, Nirmalathas A. Best optimizer selection for predicting bushfire occurrences using deep learning. Natural Hazards. 2020 May 29;