Machine Learning-based Liver Cancer Classification using Gene Expression Microarray Data

1st Amena Mahmoud
Dept. of Information and Communication Sciences, Faculty of Science and Technology, Sophia University, Japan.
Amena_mahmoud@sophia.ac.jp

2nd Syeda Meraj
Kulliyyah of Information & Communication Technology, International Islamic University (IIUM), Gombak Campus, 50728, Kuala Lumpur, Malaysia
syedashaizadi@gmail.com

3rd Shilpa Saini Sapna Juneja
Chandigarh University, Mohali, India,
shilpa.saini211986@gmail.com

4th Kazim Raza Talpur
Faculty of Technology and Informatics,
Universiti Teknologi Malaysia, Kuala Lumpur
talpurkazim@gmail.com

5th Asadullah Shah
Kulliyyah of Information & Communication Technology, International Islamic University (IIUM), Gombak Campus, 50728, Kuala Lumpur, Malaysia
asadullah@iium.edu.my

6th Shilpa Saini
Chandigarh University, Mohali, India,
shilpa.saini211986@gmail.com

7th Wesam Ahmed
Department of Information Technology, Faculty of Computers and Artificial Intelligence, South Valley University, Hurghada, Egypt;
wesam.elbaz@fcih.svu.edu.eg

2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS) | 979-8-3503-6314-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETAS62372.2024.11119847

***Abstract*—Detecting a liver tumor early and accurately can save lives because the liver is an important and multifunctional human organ. Machine learning algorithms have recently emerged as effective tools for enhancing liver cancer categorization using gene expression microarray data. This study proposes a supervised machine learning-based approach for liver cancer diagnosis that influences gene expression profiles to achieve an accurate diagnosis. A large sample size is crucial to be obtained and leads to a precise and reliable outcome. In this research, we combine multiple datasets from the Curated Microarray (CuMiDa) Database with the same features and use machine-learning models. Random forest (RF) model, SVM model, Xgboost model, K-nearest neighbor (KNN) model, and Decision tree (DT) model, and are used as classification models for classifying liver cancer using gene expressions. The results indicate that effect size and classification accuracies increase, while variances in effect size shrink with the increase in sample size. The results reveal that the RF model has better accuracy of 96.55%.**

Keywords—Liver Cancer Classification; Machine learning; Gene Expression Microarray Bioinformatics.

I. INTRODUCTION

In the light of developments in genomic and molecular technology, significant progress has been made in finding and identifying potential markers that can be utilized in the detection of liver cancer. The subsequent section will review these markers, which are categorized according to the techniques and methods used in their discovery [1][2]. Early detection of liver cancer by using serological AFP tests along with ultrasound has been reported to increase the respectability of the tumor and the survival rate of the patients. However, recent studies have indicated that the sensitivity and specificity of AFP tests vary from 39-64% and 76-90%, respectively, and the effectiveness of ultrasound examination depends largely on the skill of the examiners. As a result, there is an increasing interest in finding effective methods of early detection of liver cancer [3].

Several types of cancer have been predicted such as breast cancer, lung cancer, and liver cancer using machine learning (ML), which is a subset of artificial intelligence that allows computers to learn from training data. ML has outperformed clinicians in their ability to predict malignancy. Beyond cancer, these technologies may also enhance the prognosis, quality of life, and diagnosis of patients suffering from a variety of other diseases. As a result, it is critical to develop new patient-beneficial programs and to enhance existing AI and ML technologies [4][5].Our study's primary contributions are: presenting a diagnostic framework for liver cancer, comparing the performance of machine learning classifiers, using machine learning and multi-omic data allows for a more precise classification of liver cancer patients compared to current methods.In the remainder of the paper, a discussion to the current study, describing the methodology employed in this study, including the dataset used and the architecture of the proposed learning models. We then presented and discussed the results of our experiments, followed by a comprehensive analysis of the findings, limitations, and advantages. Finally, we concluded with a discussion of the implications of our study and the future directions for research in this field.

II. RELATED WORK

Except for bioinformatic analysis, the utilization of machine-learning techniques in genomics datasets to ascertain the primary genes accountable for HCC is limited. Wang et al. [6] identified thirteen therapeutically promising, clinically significant target genes by combining actual HCC tumor expression data with aggregated, routinely spaced short palindromic repeats and employing genome-wide growth-depletion screens. By combining bioinformatics analysis with machine-learning algorithms, it is therefore highly probable that significant target genes can be identified.For the detection of liver cancer, Lailil et al. [7] introduced a system utilizing machine learning techniques such as C5.0 Decision Tree, SVM, KNN, Naïve Bayes, and

979-8-3503-6314-2 /23/$31.00 ©2024 IEEE

feature selection method which is based on the entropy value to select the significant gene expression. In terms of accuracy, the experimental results of C5.0 Decision Tree accuracy (94%) and the computation time for algorithms using feature selection for liver cancer prediction is much shorter than without feature selection. As an approach to liver cancer diagnosis, Zhi et al.[8] predicted treatment response byusing the selection operator (LASSO), and least absolute shrinkage algorithm after the first TACE procedure[9-11]. For the construction of prediction models, six machine learning algorithms were employed, including XGBoost, decision trees, support vector machines, random forests, and complete convolutional neural networks[12]. In terms of accuracy, the RF model performed the best (78%).

III. METHODOLOGY

We describe our approach, our dataset, our preprocessing steps, and the techniques we used to diagnose liver cancer in this section. Fig. 1 illustrates our proposed framework.

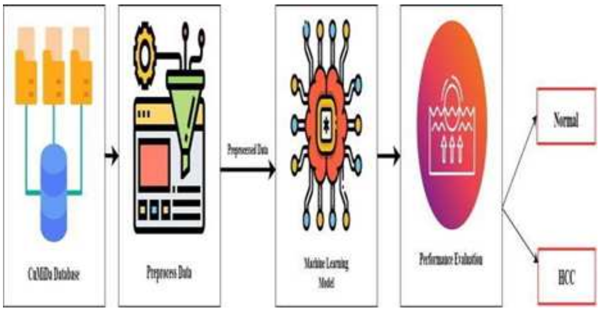

Fig. 1. Framework of the proposed work

Our work uses the CuMiDa (Curated Microarray Database) [9], a collection of 78 carefully selected microarray data sets for Homo sapiens, all of which have been carefully examined using rigorous filtering criteria based on 30,000 microarray experiments collected from Gene Expression Omnibus. Sample quality analysis, background correction, normalization, and manual editing of all data sets were applied individually [13-15]. It provides tools for users to search for and download experiments as well as curated gene expression profiles [16]. The two types of liver cancer are normal and Hepatocellular carcinoma (HCC) [17]. There are no missing values in this dataset. A total of 236 normal and 245 HCC are included in this dataset. A pre-processing operation is performed on the data to ensure the best possible result. The type column is converted to numerical data 0, and 1, where 0 represents HCC and 1 represents normal [18-19]. After that, we used the Standard Scaler which each feature's distribution is transformed to have a mean of zero and a standard deviation of one based on the principle of normalization [20]. The process prevents any one feature from dominating the learning process due to its greater magnitude by setting them all on the same scale. Additionally, Table 1 illustrates how the dataset attributes were statistically computed. These calculations may include means, standard deviations, minimums, maximums, and quartiles, which provide information about how the data is distributed and how it behaves. The dataset is balanced as shown in Figure 2.

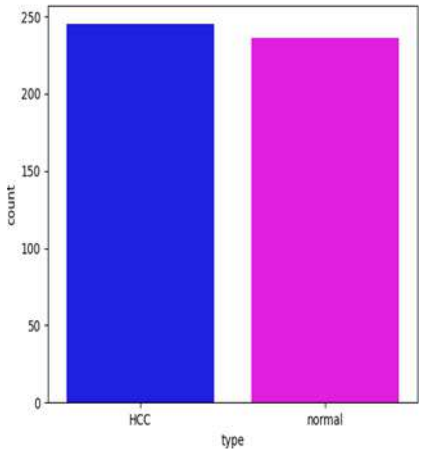

Fig.2. The types of the dataset

When an algorithm does supervised learning, it uses labeled training data to learn and estimate the results of unanticipated inputs [10-15].

IV. RESULTS AND ANALYSIS

In this proposed research, supervised learning approaches of ML for the classification of liver cancer types based on gene expressions dataset were applied. We have tested several algorithms, and the data has been split into a training set (70%) and a testing set (30%). The experiments are carried out using Google Colab and the ML algorithms were implemented using Python and Scikit-Learn [21- 23]. In Table 1, we compare the performance metrics of all proposed classifiers after feature optimization. In Figure 2, the accuracy of the RF model, SVM model, DT model, XGBoost model, and KNN model is compared. According to Table 1 and Figure 2, Based on the comparison analysis, the RF classifier has the highest accuracy, 96.55% (highlighted in bold), and is more accurate than any other research work listed. It achieves an accuracy of 96.55%, F1 score of 96.29 %, recall of 98.48 %, precision of 94.2 %, and MCC of 93.15 %. The RF model's impressive performance can be attributed to its ability to reduce overfitting which improves decision tree accuracy. The confusion matrix was used to compare each class with the other and observe how many samples were misclassified. Figure 3 shows the performance of these models using the confusion matrix. Figure 4 illustrates the AUC-ROC curve which is an indicator of how well our machine learning classifier is performing. AUC= 1 indicates that the RF classifier can correctly distinguish between the class points[24-25].

TABLE I. RESULTS OF DIFFERENT CLASSIFICATION APPROACHES

Model	Accuracy	Precision	Recall	F1-score	MCC
KNN	88.96	84.72	92.42	88.40	78.18
SVM	79.31	95	57.57	71.69	61.32
XGBoost	93.79	91.3	95.45	93.33	87.6
Decision Tree	84.13	80.28	86.36	83.21	68.38
Random Forest Classifier	96.55	94.2	98.48	96.29	93.154

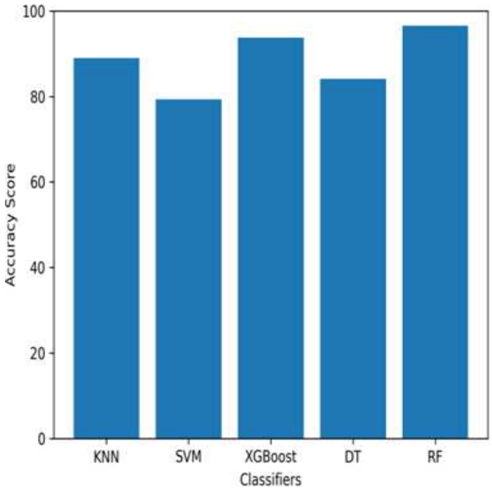

Fig. 3 Comparison between models in terms of accuracy

A. *Significance of the findings of our proposed models*

Our findings have important implications for cancer diagnosis, prognosis, and treatment. Cancer can be more precisely diagnosed and treated with improved outcomes with the application of machine learning techniques. Machine learning can analyze massive amounts of data rapidly and accurately, giving it a significant advantage over conventional approaches, and it may discover small changes between cancer kinds, subtypes, and stages that humans maymiss. In cancer classification, machine learning can also improve patient stratification, which allows inicians to provide individual treatment plans. his can decrease treatment-related adverse ects and increase treatment effectiveness.

B. *Limitations of the findings of our proposed models*

We found that ML algorithms are capable of classifying cancer into normal and HCC. However, the limitations of the study must be acknowledged. The problem is that although we evaluated and compared only five ML algorithms, many other algorithms weren't considered.

V. Conclusion

The term "Cancer" refers to a set of diseases caused by abnormal cellular growth that can spread throughout the body. The disease of cancer is predicted to become the deadliest in the future, so early diagnosis and identification are essential to its control. This paper uses the most widely used supervised machine learning which has made significant progress in the detection of cancer such as RF, DT, KNN, and XGBoost model. The first step involves collecting a comprehensive data set of gene expression microarray data from liver cancer patients and preprocessing the data to ensure its quality and compatibility. The performance of the classification models was assessed by calculating six assessment metrics. The study shows that machine learning techniques are effective at detecting disease automatically with high accuracy. In the future, we will use graphs and transformer networks to predict liver cancer disease.

References

[1] C. Venkatesan, D. Balamurugan, et al., "Efficient machine learning technique for tumor classification based on gene expression data," in 2022 8th ..., 2022.

[2] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," BMC medical genomics, 2020.

[3] J. Racle and D. Gfeller, "EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data," in Methods for Cancer Immunotherapy: Methods and Protocols, 2020, Springer.

[4] T. S. Lim and J. K. Kim, "Is liver biopsy still useful in the era of non-invasive tests?," Clinical and Molecular Hepatology, 2020.

[5] J. Neuberger and O. Cain, "The need for alternatives to liver biopsies: non-invasive analytics and diagnostics," Hepatic Medicine: Evidence and Research, 2021.

[6] G. Bergers and S. M. Fendt, "The metabolism of cancer cells during metastasis," Nature Reviews Cancer, 2021.

[7] X. Jin, Z. Demere, K. Nair, A. Ali, G. B. Ferraro,T. Natoli, et al., "A metastasis map of human cancer cell lines," Nature, 2020.

[8] W. Wang and C. Wei, "Advances in the early diagnosis of hepatocellular carcinoma," Genes & diseases, 2020.

[9] Feltes, B.C.; Chandelier, E.B.; Grisci, B.I.; Dorn, M. CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. Journal of Computational Biology, 2019.

[10] Das A., Acharya U.R., Panda S.S., Sabut S. Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. Cogn. Syst. Res. 2019;54:165–175. doi: 10.1016/j.cogsys.2018.12.009.

[11] Ghoniem R.M. A novel bio-inspired deep learning approach for liver cancer diagnosis. Information. 2020;11:80. doi: 10.3390/info11020080.

[12] Dong X., Zhou Y., Wang L., Peng J., Lou Y., Fan Y. Liver cancer detection using hybridized fully convolutional neural network based on deep learning framework. IEEE Access. 2020;8:129889–129898. doi: 10.1109/ACCESS.2020.3006362.

[13] Sureshkumar V., Chandrasekar V., Venkatesan R., Prasad R.K. Improved performance accuracy in detecting tumor in liver using deep learning techniques. J. Ambient. Intell. Humaniz. Comput. 2021;12:5763–5770. doi: 10.1007/s12652-020-02107-7.

[14] Kaur A., Chauhan A.P.S., Aggarwal A.K. An automated slice sorting technique for multi-slice computed tomography liver cancer images using convolutional network. Expert Syst. Appl. 2021;186:115686. doi:10.1016/j.eswa.2021.115686.

[15] Shukla P.K., Zakariah M., Hatamleh W.A., Tarazi H., Tiwari B. AI-DRIVEN novel approach for liver cancer screening and prediction using cascaded fully convolutional neural network. J. Healthc. Eng. 2022; 2022:4277436. doi: 10.1155/2022/4277436.

[16] J. Lobo, R. Ohashi, B. M. Helmchen, N. J. Rupp et al., "The morphological spectrum of papillary renal cell carcinoma and prevalence of provisional/emerging renal tumor entities with papillary growth," Biomedicines, vol. 2021.

[17] Kanwal, Samina, et al. "Feature Selection for Lung and Breast Cancer Disease Prediction Using Machine Learning Techniques." 2022 1st IEEE International Conference on Industrial Electronics: Developments & Applications (ICIDeA). IEEE, 2022.

[18] Sharma, Sarang, et al. "Transfer learning-based modified inception model for the diagnosis of Alzheimer's disease." Frontiers in Computational Neuroscience 16 (2022): 1000435.

[19] Anand, Vatsala, et al. "Weighted average ensemble deep learning model for stratification of brain tumor in MRI images." Diagnostics 13.7 (2023): 1320.

[20] Dhiman, Gaurav, et al. "A novel machine- learning-based hybrid CNN model for tumor identification in medical image processing." Sustainability 14.3 (2022): 1447.

[21] Kaur, Gurjinder, et al. "Automatic Identification of Glomerular in Whole-Slide Images Using a Modified UNet Model." Diagnostics 13.19 (2023): 3152.

[22] Sharma, Neha, et al. "UMobileNetV2 model for semantic segmentation of gastrointestinal tract in MRI scans." Plos one 19.5 (2024): e0302880.

[23] Sharma, Sandhya, et al. "Performance evaluation of the deep learning based convolutional neural network approach for the recognition of chest X- ray images." Frontiers in oncology 12 (2022): 932496.

[24] Hossain, M. Shamim, and Ghulam Muhammad. "Deep learning based pathology detection for smart connected healthcare." IEEE Network 34.6 (2020): 120-125.

[25] Masud, Mehedi, et al. "Pre-trained convolutional neural networks for breast cancer detection using ultrasound images." ACM Transactions on Internet Technology (TOIT) 21.4 (2021): 1-17.