Application of Machine Learning for Diagnosis of Liver Cancer

Kushagra Baser
Dept. of Electronics and Communication Engineering, Nirma University
Ahmedabad, India
20bec058@nirmauni.ac.in

Maurya Shah
Dept. of Electronics and Communication Engineering Nirma University
Ahmedabad, India
20bec065@nirmauni.ac.in

***Abstract*— A fundamental obstacle in cancer research is predicting the course of liver cancer disease. In this study, we examined models for projecting the progression of liver cancer known as hepatocellular carcinoma (HCC). Technology advancement has led to the production of patient data in various disease phases as a result of its prevalence as a major cause of liver cancer mortality. The ability of machine learning algorithms to predict HCC liver cancer using a publically available dataset would need to be evaluated. We advise using machine learning techniques such as Naive Bayes, Neural Networks, SVM, SGD, and Logistic Regression. Using actual medical data on liver cancer, we determine their area under the curve and evaluate its correctness. Experimental results show that neural networks perform better.**

Keywords*— *Machine Learning, Ordinal, Nominal, Confusion Matrix, Cancer Genomics, Medical Diagnosis, Hepatocellular Carcinoma (HCC)

I. INTRODUCTION

By 2022, 17 million people may have received a liver cancer diagnosis[1]. Healthcare professionals having access to effective ways of spotting cancer occurrences will greatly enhance the medical diagnosis of cancer patients. To do this, professionals from a variety of sectors worked together to develop novel computational techniques that improve medical diagnoses. A computer algorithm based on gene expression data was proposed by Stokowy et al. [2] to locate thyroid nodules prior to surgery. An "Extreme Learning Machine (ELM)" was proposed by Zhang et al. [3] for lung, lymphoma, and other microarray cancer data sets.

According to experimental findings that used accuracy as a metric, ELM outperformed other types of support vector machines. Wang et al.'s machine learning proposal[4] centred on the classification of tumours as a problem. With the aid of microarray data, the novel method was contrasted with conventional methods. In order to more accurately diagnose cancer, modern machine learning algorithms have been developed.

Automatic learning and optimal prediction are the focus of the computer science discipline known as machine learning, or ML. Many professionals categorise machine learning into three separate phases:

1) Educating oneself by observation and experience.
2) Keep doing activities like categorization, prediction, etc.
3) Improve output by learning from both historical and current data.

Although earlier technologies have advanced cancer diagnosis, their forecast accuracy depends heavily on computational techniques. In view of the recent discovery of computational approaches to improve performance in the field of machine learning[5], [6], using new machine learning algorithms to clinical data acquired using a variety of technologies is crucial.
The remainder of the paper is divided into the following sections. The work of numerous authors in the same topic is reviewed in Section II. The dataset and study approach are explained in Section III. In Section IV, experiments and findings utilising ML algorithms on real-world liver cancer-related data are presented. Section V of this article deals with predicting one year survival of patients with Hepatocellular Carcinoma using a Logistic Regression Model. Section VI of this article deals with findings and analysis.Section VII of this article concludes by highlighting a number of areas for additional research.

II. LITERATURE REVIEW

The Kumardeep Chaudhary et al. [7] deep learning (DL)-based model can accurately identify between patient subpopulations with various survival odds across six distinct cohorts. A DL-based, survival-sensitive model that predicts outcome as well as a model that takes into account both genomic and clinical factors were built using data from the "Cancer Genome Atlas (TCGA)", including "RNA sequencing (RNA-Seq)", "microRNA sequencing (miRNA-Seq)", and "Methylation profiles" for 360 patients with hepatocellular carcinoma (HCC).

Amit Das et al. [8] presented a unique method called watershed Gaussian-based deep learning (WGDL) for effectively highlighting liver cancer lesions in CT data. After 200 iterations, the DNN classifier employed in this study achieved a 99.38% classification accuracy and a 98.18% Jaccard index with a relatively tiny validation loss of 0.062.

Using 31,608 pictures from 1,210 liver cancer patients, Shi-hui Zhen et al. [9] trained CNNs to provide seven- way, binary, and three-way malignancy classifiers. The models were validated by an extensive, varied, and independent external cohort of 201 patients(6,816 pictures). Even without image improvements, CNN can discriminate between malignant and benign liver tumors(AUC 95%). Using a unique convolutional neural network (CNN) that depended on unprocessed images and clinical data, classification of HCC and metastatic tumors was improved.

Charlie A. Hamm et al. [10] created a distinctive CNN by iterative network construction and training instance optimisation. Included are three convolutional layers, two layers with maximum pooling, and two layers that are fully connected. 60 test subjects and 434 hepatic lesions from 6 categories were used for training. The results of the trial showed 92% success, 92% sensitivity, and 98% specificity. Sn and Sp averaged 90% and 98% in a single run of random, undetected occurrences.

III. MATERIALS AND METHODS

A. *Data Set*

Data set of Liver Cancer patients[11] has been used in this study. Dataset contains 583 records of patients with 10 attributes required to detect liver cancer . Another Dataset contains record of 165 patients with 50 attributes.

B. *Methodology*

The flow of implementation is depicted in fig 1.

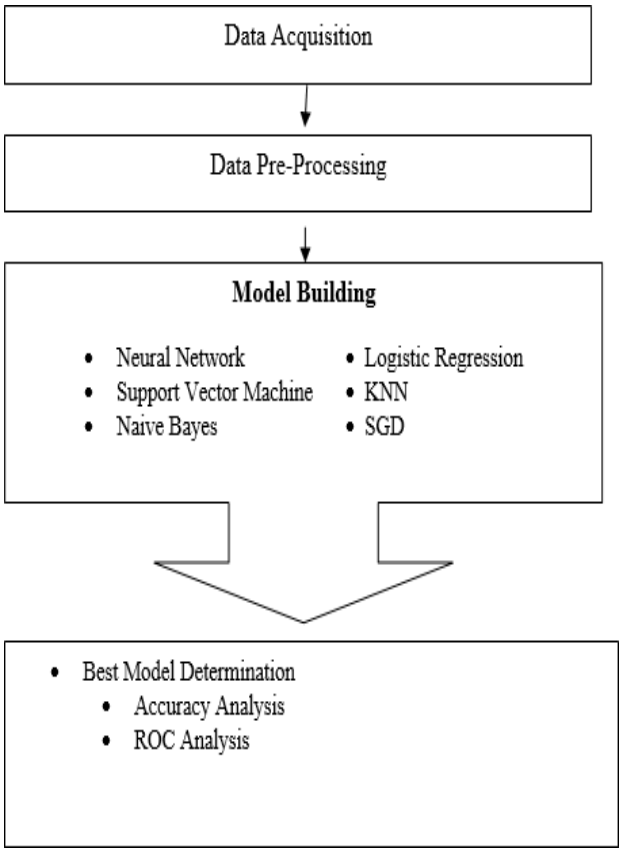

Fig. 1. Proposed Methodology

1) Dataset Splitting

The competence of a machine-learning model has been evaluated using a cross-validation procedure that is 10 fold.

2) Model Training

Following algorithms have been used to train the model in this study.

a) Artificial neural network

Artificial neurons, which are basic mathematical operations, serve as the building blocks of every artificial neural network (ANN). These models use mathematical operations including addition, multiplication, and activation.It is proof that a synthetic neuron modifies its multiplier and assigns a weight to its inputs[12]. The weighted biases and all inputs are combined at the centre of the artificial neuron. The activation of weighted inputs and biases (transfer function) can produce synthetic neurons. There are several different ANN structures. MLP is a directed graph having multiple levels in particular.

The data from the layer below is processed by nodes in one layer, who then compute a function value and transmit it to the node in the subsequent layer. The names of the input, hidden, and output layers are all distinct. The layers between the input and the output are referred to as the "hidden layer". A weighted sum of their inputs activates the hidden and output layers. Among the many uses for neural networks, prediction, financial modelling, and medical bioinformatics are only a few[13].

b) *Support* vector machine(SVM) classification

Support vector machines (SVMs) are used in machine learning for prediction and classification. These tactics aim for an immediate resolution while avoiding current problems. To avoid overfitting, SVM expands structural risk reduction. A good extrication hyperplane is desired for SVM classifiers. Outstanding statistical characteristics are present in an ideal unscrambling hyperactive plane. SVC, on the other hand, is defined first. The kernel technique for nonlinear decision planes broadens the scope of SVC. Noisy data can be controlled via slack variables[14].

c) *Naive* Bayes classifier

When it comes to operational classifiers, the Bayesian network is a serious candidate. These networks' conditional probabilities are built on a structure resembling a network. In a BN diagram, the provincial variables and their relationships are represented by the nodes and edges. A dataset needs to be gently cleaned and thoroughly labelled before using a Naive Bayes classifier[15], [16].

d) *Logistic* regression classifier

When given a real-valued input vector, logistic regression may carry out discriminative classifications. An input vector measurement known as a feature serves as a data predictor. Data with multiclass categories can also use logistic regression. For instance, logistic regression[17] is one technique for determining the probability P of a dichotomous result in a Bernoulli trial.

e) k-*nearest* neighbour

K-nearest-neighbor (kNN) is a training method for instance-based learning that does not require a model. It switches to a simpler approach and looks for a class that possesses some of the necessary characteristics for the input domain. The kNN algorithms classify the data set by using the most prevalent symptoms among a given group of cases. However, the requirement of the number k[18, [19]] limits the scope of this categorization assignment.

f) *Stochastic* gradient descent(SGD)

SGD iteratively improves the smoothness of an objective function (such as a differentiable or subdifferentiable one). If you don't have the entire set of data, you can optimise using an approximation of the gradient (derived from a small group of randomly chosen data). Naturally, faster iterations lead to slower convergence rates, but this is particularly true for high-dimensional optimisation problems[20].

g) *Multi layer Pereceptron(MLP)*

The feed forward neural network is supplemented by the multi layer perceptron (MLP). The input layer, output layer, and hidden layer are the three different kinds of layers that make it up. The input layer is where the input signal for processing is received. The output layer completes the desired task, such as prediction and classification.

3) Classification performance measurement

The categorization of liver cancer samples was performed using four supervised machine learning algorithms, and the accuracy of the classification was assessed using ten rounds of cross-validation. They evaluated the categorization models using eight distinct quality indices. We evaluated the performance of the categorization analyses using these quality markers. For instance, samples without liver cancer disease were categorised as "positive," whereas samples with liver cancer disease were categorised as "negative."

a) Performance Evaluation Metrics

These measures are used to assess the model's efficacy in terms of its accuracy[21]. Moreover, we utilized the assessment according to the aforementioned criteria to make sure the confusion matrix didn't deceive us.

IV. FINDINGS AND ANALYSIS

Based on the 10 attributes for 583 patients detailed in the study's methodology and materials, this report compares three machine- learning methods for liver prediction. In cross-validation, data samples were split into twenty parts and used for testing andtraining.

Figures 2 and 3 depicts the confusion metrices for several prediction models, including NN, logistic regression.

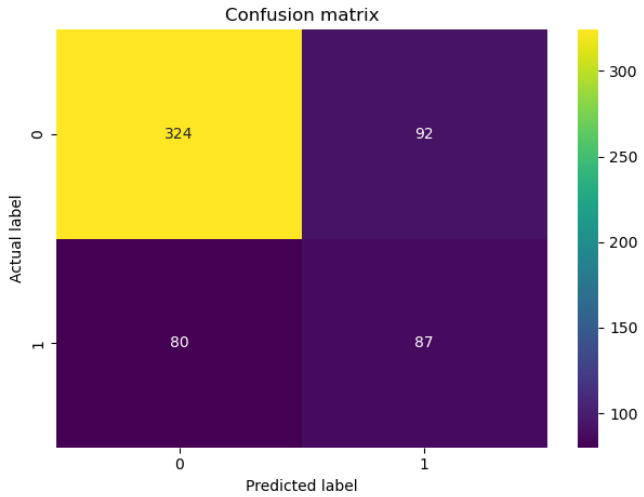

Fig.2. Confusion matrix for Neural Network

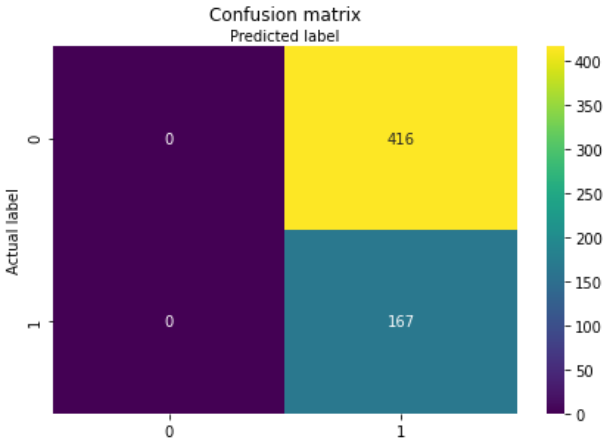

Fig.3. Confusion Matrix for Regression Technique.

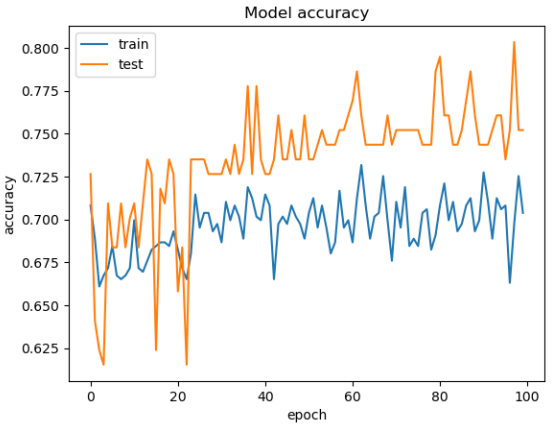

Fig.4. Model accuracy for Neural Network..

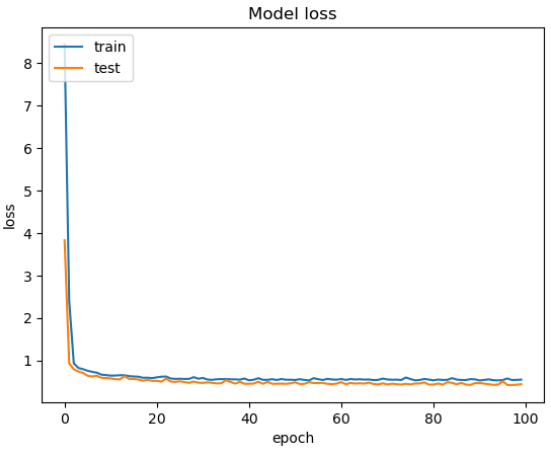

Fig.5. Model Loss for neural Network.

```
The best accuracy is at
solver                  lbfgs
activation_function      tanh
hidden_layers          (5, 2)
accuracy             0.818182
loss                 0.358918
Name: 18, dtype: object
```

Fig.6. Model accuracy for MLP

V. PREDICTING ONE YEAR SURVIVAL OF PATIENTS WITH HEPTOCELLULAR CARCINOMA : LOGISTIC REGRESSION MODEL

This model's goal is to lower the death rate related to hepatocellular carcinoma by determining whether a patient with a diagnosis of hepatocellular carcinoma would survive through one year in conjunction with existing guidelines.

Clinicians, patients, and policymakers alike would find this model to be very helpful as they consider various cancer treatment options and decision-making processes. Machine learning and statistical models will ideally find use in presenting a more comprehensive and all-encompassing patient profile.

The patient data for 165 former patients of the Hospital and University Centre of Coimbra (Portugal) make up this dataset on hepatocellular carcinoma. The dataset includes 49 features that were chosen in accordance with the EASL-EORTC Clinical Practise Guidelines (European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer). The "Class" target variable, which is expressed as 0 (dead) and 1 (living), represents each patient's survival at one year.

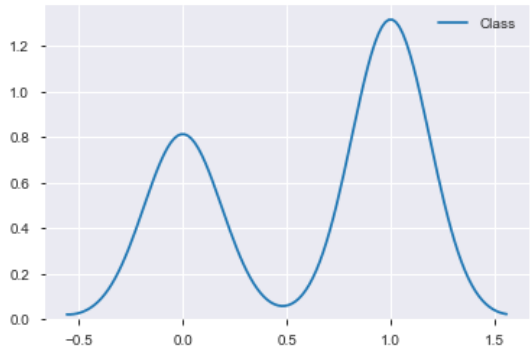

Fig.7. Distribution Of Target Class : Alive or Dead.

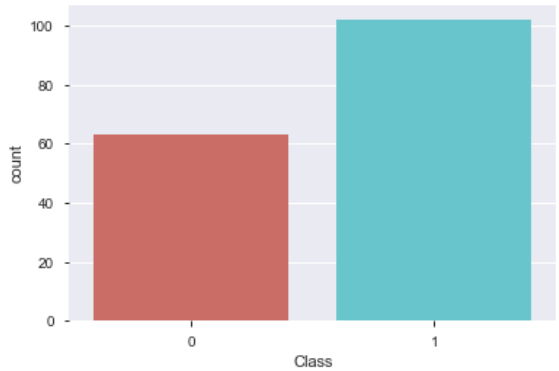

Fig.8. Distribution of Trget Class.

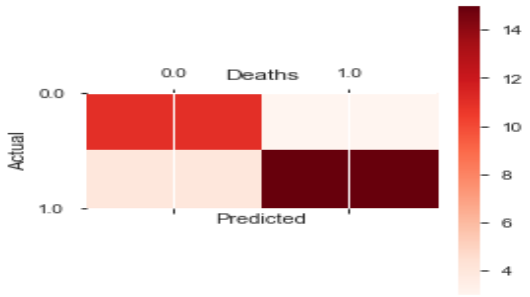

Fig.9. Confusion Matrix Graph

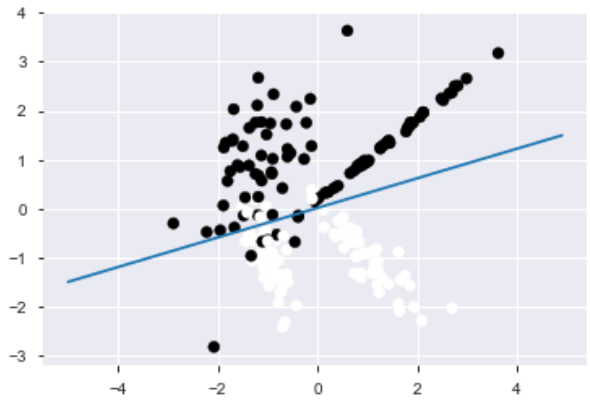

Fig.10 Decision Boundary Graph For Logistic Regression.

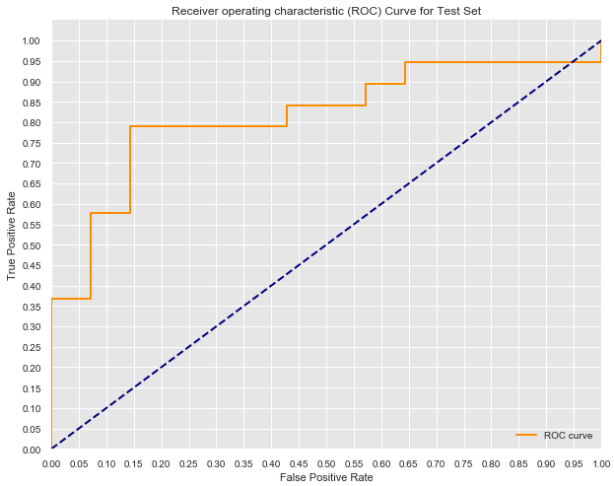

Fig.11. AUC Graph/ROC Curve .

VI. Finding and Analysis

MODEL	ACCURACY
1. Neural Network	70.49%
2. Regression	40.14%
3. Multi Layer Perceptron	81.81%

When compared to other machine learning strategies, Multi Layer Perceptron performed best with a maximum classification accuracy(CA) of 81.81%.

For HCC using Logistic regression model the AUC was 0.8157.

Using the Cconfusion metrices, researchers were able to determine which studies produced the most accurate positive results.

VII. CONCLUSION

In order to preserve lives and influence therapy, a liver cancer diagnosis is crucial. The findings of this study demonstrate that liver cancer can be identified using machine learning techniques. The accuracy of our MLP that was cross-validated was 81.81%, and the accuracy of our other classifiers was similarly at a usable level. In this work, a classification method was employed to make liver cancer predictions based on a variety of distinct variables. We intend to extract more attributes from the present dataset in the future to increase the accuracy attained. We will also employ a wider variety of ML and deep learning techniques, including transfer and deep learning, for the best results.

A clustering oversampling technique was used to capture the variability for each patient profile as well as the dataset class imbalance. In the end, a logistic regression model was employed since it had a higher predictive value than other models like SVM and random forest.

Using the performance metrics of accuracy, AUC, and F-1 score, the logistic model's performance was evaluated. The false positive rate was kept as low as possible because it is quite easy to forecast mistakenly that a patient will live at least one more year (Class=1).

On the testing dataset, the logistic regression model produced results with an accuracy of 82%, an AUC of 83%, and an F-1 score of 83%. The estimated false positive rate was 14%. With the addition of the SMOTE oversampling approach to address class imbalance and the development of two new features, MELD score and FIB-1 Index score, all performance metrics improved.

This approach can support patient decision-making by offering a practical HCC diagnostic tool. More reliable and effective healthcare judgements can be made earlier in the illness lifecycle by combining machine learning's capacity to analyse massive datasets in a short amount of time with decision makers' contextual expertise.

Future work will focus on developing predictive models using larger patient datasets, a challenge in the healthcare industry that is easier said than done. Additionally, using these models for different illness classification scenarios may have a positive influence on decision-makers' approaches to diagnoses and the best course of therapy. Future healthcare models that attempt to maintain patient profile heterogeneity may find promise in the idea of using clustering techniques to fill in patients' missing data.

REFERENCES:

[1] J. Huang et al., “Worldwide Burden, Risk Factors, and Temporal Trends of Ovarian Cancer: A Global Study,” Cancers 2022, Vol. 14, Page 2230, vol. 14, no. 9, p. 2230, Apr. 2022, doi: 10.3390/CANCERS14092230.

[2] T. Stokowy et al., “A two miRNA classifier differentiates follicular thyroid carcinomas from follicular thyroid adenomas,” Mol Cell Endocrinol, vol. 399, pp. 43–49, Jan. 2015, doi: 10.1016/J.MCE.2014.09.017.

[3] R. Zhang, G. bin Huang, N. Sundararajan, and P. Saratchandran, “Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis,” IEEE/ACM Trans Comput Biol Bioinform, vol. 4, no. 3, pp. 485–494, Jul. 2007, doi: 10.1109/TCBB.2007.1012.

[4] Y. Wang et al., “Gene selection from microarray data for cancer classification—a machine learning approach,” Comput Biol Chem, vol. 29, no. 1, pp. 37–46, Feb. 2005, doi: 10.1016/J.COMPBIOLCHEM.2004.11.001.

[5] R. G. Tiwari, M. Husain, B. Gupta, and A. Agrawal, “Amalgamating contextual information into recommender system,” 2010. doi: 10.1109/ICETET.2010.110.

[6] N. Kumar Trivedi, S. Simaiya, S. Kumar Sharma, and U. Kumar Lilhore, “Machine learning View project credit card fraud detection using machine learning View project COVID-19 Pandemic: Role of Machine Learning & Deep Learning Methods in Diagnosis,” Int J Cur Res Rev |, vol. 13, 2021, doi: 10.31782/IJCRR.2021.SP192.

[7] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, “Deep learning–based multi-omics integration robustly predicts survival in liver cancer,” Clinical Cancer Research, vol. 24, no. 6, pp. 1248–1259, Mar. 2018, doi: 10.1158/1078-0432.CCR-17- 0853/116627/AM/DEEP-LEARNING-BASED-MULTI-OMICSINTEGRATION.

[8] A. Das, U. R. Acharya, S. S. Panda, and S. Sabut, “Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques,” Cogn Syst Res, vol. 54, pp. 165–175, May 2019, doi: 10.1016/J.COGSYS.2018.12.009.

[9] S. H. Zhen et al., “Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data,” Front Oncol, vol. 10, p. 680, May 2020, doi: 10.3389/FONC.2020.00680/BIBTEX.

[10] C. A. Hamm et al., “Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI,” European Radiology 2019 29:7, vol. 29, no. 7, pp. 3338–3347, Apr. 2019, doi: 10.1007/S00330-019-06205-9.

[11] “COVID-19 effect on Liver Cancer: Introduction | Kaggle.” https://www.kaggle.com/code/fedesoriano/covid-19-effect-onliver-cancer-introduction/data (accessed Oct. 15, 2022).

[12] A. Manoharan, K. M. Begam, V. R. Aparow, and D. Sooriamoorthy, “Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review,” J Energy Storage, vol. 55, p. 105384, Nov. 2022, doi: 10.1016/J.EST.2022.105384.

[13] R. G. Tiwari, A. Misra, V. Khullar, A. K. Agarwal, S. Gupta, and A. P. Srivastava, “Identifying Microscopic Augmented Images using Pre-Trained Deep Convolutional Neural Networks,” 2021. doi: 10.1109/ICTAI53825.2021.9673472.

[14] P. Chhajer, M. Shah, and A. Kshirsagar, “The applications of artificial neural networks, support vector machines, and long– short term memory for stock market

prediction," Decision Analytics Journal, vol. 2, p. 100015, Mar. 2022, doi: 10.1016/J.DAJOUR.2021.100015.

[15] X. Feng, S. Li, C. Yuan, P. Zeng, and Y. Sun, "Prediction of Slope Stability using Naive Bayes Classifier," KSCE Journal of Civil Engineering 2018 22:3, vol. 22, no. 3, pp. 941–950, Mar. 2018, doi: 10.1007/S12205-018-1337-3.

[16] F. J. Yang, "An implementation of naive bayes classifier," Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018, pp. 301–306, Dec. 2018, doi: 10.1109/CSCI46756.2018.00065.

[17] M. Goswami and N. J. Sebastian, "Performance Analysis of Logistic Regression, KNN, SVM, Naïve Bayes Classifier for Healthcare Application During COVID-19," Lecture Notes on Data Engineering and Communications Technologies, vol. 96, pp. 645–658, 2022, doi: 10.1007/978-981-16-7167-8_47/COVER.

[18] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 1–11, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[19] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," ACM Computing Surveys (CSUR), vol. 54, no. 6, Jul. 2021, doi: 10.1145/3459665.

[20] N. Ketkar, "Stochastic Gradient Descent," Deep Learning with Python, Apress, Berkeley, CA., pp. 113–132, 2017, doi: 10.1007/978-1-4842-2766-4_8.

[21] R. Yacouby Amazon Alexa and D. Axman Amazon Alexa, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," pp. 79–91, Nov. 2020, doi: 10.18653/V1/2020.EVAL4NLP-1.9.