International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

A Model to Predict Survival among Pancreatic Cancer Patients using Supervised Machine Learning

Simbarashe N. Manyetu[1], Monika Gondo[2]

[1]Computer Science Department, Harare Institute of Technology
manyetu[at]gmail.com

[2]Computer Science Department, Harare Institute of Technology
monicagondo[at]gmail.com

Abstract: ***Cancer is a diverse illness with uncontrolled cell division and the propensity to spread. It has continuous cell division, no programmed cell death, tissue invasion, and blood vessel formation at its location. Oncologists can use cancer survival rates to determine a patient's prognosis and treatment approach. Age, race, cancer site, stage, grade, number of tumors, kind of cancer; and other factors affect patient survival. This study focused on Pancreatic Cancer patient survival prediction using CNN on a dataset that with 7558 labelled images.The model predicted cancer survival with a maximum score of 1 which is the highest possible score on all the metrics used. However, the confusion matrix of the model indicates that the model made 10 false positives and 9 false negatives making a total of 19 false predictions.The model's training and validation loss were 0.02 and 0.03. These loss values are low and converge to zero, indicating effective model prediction and dependability.***

Keywords: Pancreatic Cancer, Supervised Machine Learning

1. Introduction

Cancer is a heterogeneous disease with uncontrolled cell division and the capability to spread to the surrounding tissue or body, there are more than 200 different types of cancers currently detected in humans [1]. It has the characteristics like continuous cell division, no programmed cell death, invasion of tissues, and promotion of blood vessel growth at the site of occurrence. On the other hand, cancer survivability can be defined as the time period a patient remains alive after being diagnosed [2]. Cancer survival rates can be helpful for oncologists and other medical personnel to understand the prognosis and develop a treatment plan for any particular patient. Survival of a patient can be determined by the conditions like age, race, site of cancer, stage of cancer, grade of cancer, number of tumors, type of cancer, and other conditions [3].

2. Background to the study

According to 2017 figures from the American Cancer Society (ACS), the 5-year survival rate for all stages of PC is 8.5 percent [1]. Patients with an early-stage diagnosis can have a 5-year survival rate of up to 20%. Only around a quarter of patients (15%) have a surgically respectable illness when they are diagnosed [4]. Furthermore, due to the lack of effective screening technologies, the lack of sensitive and specific biomarkers, and the low frequency of PC, identifying persons at high risk or with early-stage disease is difficult which makes the traditional detection methods ineffective [5].

Early identification of cancer improves the odds of survival. Some cancers, such as Pancreatic Cancer, are difficult to identify or detect early, and the stages develop quickly [6]. This research discusses cutting-edge cancer survival prediction approaches, as well as how these techniques can be applied to predict overall survival in patients with Pancreatic Cancer. Recent studies emphasize the necessity of machine learning (ML) algorithms like support vector machines and convolutional neural networks because of the perplexing and large amounts of data. According to studies, the survival rate for Pancreatic Cancer is 41.7 percent after one year, 8.7 percent after three years, and 1.9 percent after five years [2]. However, there is no statistically significant link between illness stages and overall survival rates.

Machine learning can help us understand cancer's progression. Machine-learning techniques need validation for routine clinical use. These methods classify, forecast, and estimate. Pathologists can utilize precise predictions to learn about a patient's condition, surgical treatment, best resource usage, personalized therapy, medications to give, and better patient care to increase survival prospects [7]. Numerous studies have identified and validated promising PC biomarkers. Doppler ultrasound (DU), endoscopic ultrasound (EUS), magnetic resonance imaging (MRI), computed tomography (CT) scan, or positron emission tomography (PET) can detect pre-cancerous abnormalities in the pancreas in high-risk people but does not predict survival in identified patients[8].

Because of its potential for diagnostic and prognostic applications, the artificial neural network (ANN), which is based on the neuronal structure of the brain [4], has piqued the interest of scientists all over the world in the field of medicine [3]. It has been used in the diagnosis of heart disease [5], headache prediction, pre-diagnosis of hypertension [2], kidney stone diseases [5], classifying breast masses to identify breast cancer [6], and dermatologist-level classification of skin diseases/cancer. For PC survival rate prognosis, predictive approach models such as ML (statistical multivariate regression) and deep learning (DL) can be employed. For optimum predictive performance, the techniques must be reinforced. This proposed research, therefore, will employ supervised

machine learning in predicting Pancreatic Cancer patients' survival using knowledge of existing data.

3. Problem Statement

Cancer has a major impact on society and accounts for a significant proportion of mortalities globally. Some of the mortalities caused by cancer can be attributed to the inefficiencies of the current technologies in giving indicators and warnings which can be used to tell the survival chances of patients. These challenges have been coupled with an increase in the quantity of cancer data which has made it difficult for medical specialists to process effectively and derive meaningful results in support of patient survival analytics efforts. This study, therefore, develops and implements a supervised machine-learning model that predicts patient survival among PC patients.

4. Literature Review

a) Survivability Prediction Studies and Prediction Using Breast Cancer Data Analysis [9]

The study created an analytical technique that helps comprehend survivorship in case of missing statistics and describes patient survival characteristics. Unsupervised learning is used to sort similar data into cohorts. Patient cohort clusters were constructed using SOM and DBSCAN (DBSCAN). Clustering followed data pre-treatment. Using these clusters, supervised MLP was trained. Each of the nine clusters has a different survival time. Results indicated that SOM and DBSCAN via characteristics can explain cluster survival periods.

b) Ensemble Data Mining using SEER Data to Predict Lung Cancer Survival [10]

This study developed a prediction model for lung cancer survival. This project creates machine learning models with default settings using SEER data and WEKA. Cancer survival is determined at 6, 9, 1, 2, and 5 years. Attribute selection techniques were used to disperse 13 qualities while preserving the initial set's strength. Machine learning models include SVMs, NNs, J48, random forests, logit boost, and ensemble voting. The voting ensemble was the best lung cancer survival predictor. Voting Ensemble 11 Classifier accuracy was 73.61 for 6 months, 74.45 for 9 months, 76.80 for 1 year, 85.45 for 2 years, and 91.35 for 5 years. Based on this study, an online calculator predicts lung cancer.

c) SEER Cancer Data-Based Reproducible Survival Prediction [11]

This review evaluates past research that used SEER data to study cancer survivability to determine if the findings can be repeated. After obtaining experimental design information from 34 articles, just one yielded directly repeatable results. The article created machine learning algorithms to predict one-year and five-year cancer survival. Logistic regression and multilayer perceptrons are categorization models. To find the best cancer survival prediction model, they looked at accuracy, AUC, and f1 scores. MLP predicts better than other machine learning models.

d) Data mining techniques for predicting breast cancer survivability [16]

This article explains how machine-learning algorithms predict breast cancer survival. After data processing, 16 variables were chosen for survival prediction analysis from the SEER data set. This research used C4.5, Naive Bayes, and Artificial Neural Nets (back propagation). This article compares the construction of the dependent variable to other studies. WEKA's 10-fold cross-validation was used to generate the machine-learning models. Multilayer perceptron and Naive Bayes are poor cancer survival predictors compared to C4.5.

5. Methodology

The study employed Convolutional Neural Networks (CNN) to predict survivability among cancer patients.

a) Dataset

The study used a dataset comprising labeled pancreatic cancer images which qualified the study to be a classification problem. The dataset comprised 7 558 labelled images.

b) Machine Learning Algorithm Used

The study implemented CNN for image classification. The model is shown in Figure 1.

```
model.summary()
)

Conv1_relu (ReLU)              (None, 112, 112, 32  0           ['bn_Conv1[0][0]']
                               )

expanded_conv_depthwise (Depth (None, 112, 112, 32  288         ['Conv1_relu[0][0]']
wiseConv2D)                    )

expanded_conv_depthwise_BN (Ba (None, 112, 112, 32  128         ['expanded_conv_depthwise[0][0]']
tchNormalization)              )

expanded_conv_depthwise_relu ( (None, 112, 112, 32  0           ['expanded_conv_depthwise_BN[0][0
ReLU)                          )                                 ]']

expanded_conv_project (Conv2D) (None, 112, 112, 16  512         ['expanded_conv_depthwise_relu[0]
                               )                                 [0]']

expanded_conv_project_BN (Batc (None, 112, 112, 16  64          ['expanded_conv_project[0][0]']
hNormalization)                )
```

Figure 1: Model summary

The model used the Adam function as the activation function and evaluation was done using classification report which summarised model performance using accuracy score, F1 score, precision score and recall score. The Mean squared error and model loss were also used to evaluate the model performance.

5.1 Results of Algorithm Performance

Figure 2 depicts the outcome of CNN classifier metrics.

```
[[4032   10]
 [   9 3507]]
---------------
              precision    recall  f1-score   support

      normal       1.00      1.00      1.00      4042
     anomaly       1.00      1.00      1.00      3516

    accuracy                           1.00      7558
   macro avg       1.00      1.00      1.00      7558
weighted avg       1.00      1.00      1.00      7558
```

Figure 2: CNN Classifier Model Testing Results

The metrics show that the model predicted the cancer classes with a maximum score of 1 which is the highest possible score on all the metrics used. However, the confusion matrix of the model indicates that the model made 10 false positives and 9 false negatives making a total of 19 false predictions.

5.2 Deployment results for CNN

Deploying the model was done to test its prediction effectiveness on new data that it had not been exposed to previously. As such, the model was deployed and tested using the testing dataset which formed 20% of the original dataset. The prediction result of the deployed CNN model is shown in Figure 3.

```
predictions=cnn.predict(X_test)
predictions.reshape(-1,1)
array([[1.],
       [0.],
       [0.],
       [1.],
       [0.],
       [1.],
       [0.],
       [1.],
       [1.],
       [1.],
       [1.],
       [1.],
       [0.],
       [0.],
       [0.],
```

Figure 3: CNN prediction results on the testing data

The prediction results of the CNN deployed model shown in Figure 3 depict an extract of the head of the predicted results. The outcome also produced a singular array that with predictions that correctly reflect the classes in the dataset used. Figure 4 depicts the loss incurred by the CNN algorithm in the prediction process.

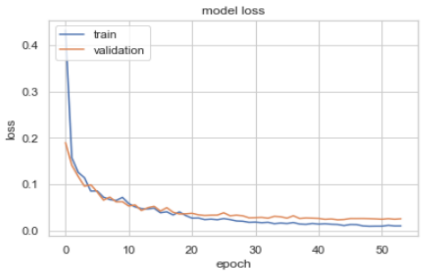

Figure 4: CNN model loss

The model loss decreased as the prediction process progressed, indicating its loss. The model's training and validation loss were 0.02 and 0.03. These loss values are low and converge to zero, indicating effective model prediction and dependability.

Figure 5 depicts the MSE and both the training and validation of the model.

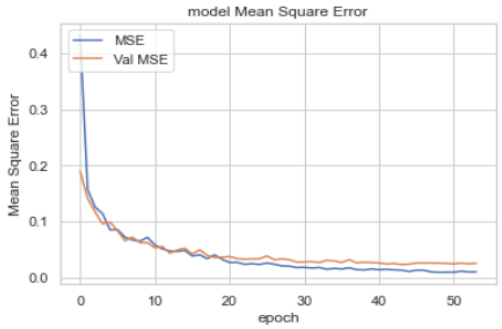

Figure 5: Model MSE

MSE model findings demonstrate low prediction error. Model error increased to 0.03 for training and 0.02 for validation. Low model errors show that it accurately predicted the survival of patients with valid scores. The confusion matrix was used to evaluate model effectiveness. The confusion matrix also known as an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one [12]. Figure 6 depicts the confusion matrix of the model.

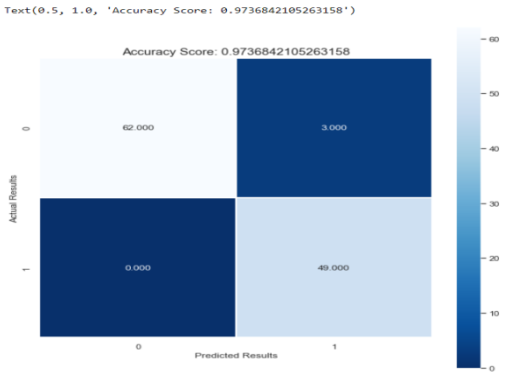

Figure 6: Confusion matrix for CNN

The model predicted successfully 62 true positives and 49 true negatives. However, the model made 3 false negative predictions in which the model classified patients in category 1 as they belong to category 0. This shows that the error level of the model is also low as indicated in the loss and MSE.

6. Conclusion

The study successfully implemented CNN in predicting cancer survival. The CNN model predicted cancer survival with a loss of 0.02 and 0.03 on both the training and validation sets. The model also predicted survival on the testing data with no false predictions. The study concludes that the implementation of CNN in pancreatic cancer patient survival is effective and produces reliable results as evidenced by the metrics.

7. Future Scope

The study focused on the implementation of CNN to predict Pancreatic Cancer patient survival. This limits the scope of the study to the application of supervised machine learning algorithms only as well as applicability to labelled datasets only. However, the changing nature of diseases makes it necessary to use unsupervised and Deep learning which also adapts to the nature of cancer conditions.

References

[1] A.-B. R, "Colon cancer survival prediction using ensemble data mining on SEER data," *In: 2013 IEEE International Conference on Cancer,* vol. 2, no. 1, pp. 23-33, 2019.

[2] F. Y, "Predictive Modeling of the Risk of Acute Kidney Injury in Critical Care: A Systematic Investigation of The Class Imbalance Problem.," *AMIA*

Summits Transl Sci Proc 2019, vol. 5, no. 3, p. 809–18, 2019.

[3] A. A, "Lung Cancer Survival Prediction using Ensemble Data Mining on Seer Data," *Sci Program 2020*, vol. 20, no. 1, p. 29–42, 2020.

[4] S. A, "Predicting breast cancer survivability using data mining techniques: In: 2010 2nd International Conference on Software Technology and Engineering (ICSTE 2020).," *IEEE 2020*, vol. 3, no. 2, pp. 10-19, 2020.

[5] M. Fotouhi S, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J Biomed Inform 2019*, vol. 12, no. 3, pp. 137-149, 2019.

[6] S. MS, "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]," *IEEE Comput Intell Mag 2018*, vol. 1, no. 3, p. 59–76.

[7] D. G, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability," *Comparison of the C4.5 and a Naive Bayes Classifie*, vol. 3, no. 2, pp. 71-79, 2019.

[8] S. A, "Treating Colon Cancer Survivability Prediction as a Classification Problem," *ADCAIJ Adv Distrib Comput Artif Intell J 2016*, vol. 5, no. 3, pp. 87-95, 2016.

[9] N. e. a. Shukla, "Breast cancer data analysis for survivability studies and prediction," *Computer methods and programs in biomedicine*, vol. 15, no. 5, pp. 199-208, 2018.

[10] R. Al-Bahrani, "Colon cancer survival prediction using ensemble data mining on SEER data. In 2013 IEEE international conference on Big Data," *IEEE*, vol. 1, no. 2, pp. (pp. 9-16), 2018.

[11] J. Ding, "Nomogram predicting the cancer-specific survival of early-onset colorectal cancer patients with synchronous liver metastasis: a population-based study," *International Journal of Colorectal Disease*, vol. 37, no. 6, pp. 1309-1319, 2022.

[12] S. Jubair, "A novel approach to identify subtype-specific network biomarkers of breast cancer survivability," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1-12, 2020.

Author Profile

Simbarashe N. Manyetu is a student at Harare Institute of Technology pursuing a Master of Technology Degree in Cloud Computing.

Monica Gondo is a lecturer in the Computer Science Department at the Harare Institute of Technology.